A Bayesian methodology for building consistent datasets for structural modeling

Applying information theory to disparate and sparse agricultural datasets for IMPACT

Daniel Mason-D'Croz[1,2], Sherman Robinson[2], Shahnila Dunston[2], and Timothy B. Sulser[2]

[1] Commonwealth Science and Industrial Research Organisation (CSIRO)
[2] International Food Policy Research Institute (IFPRI)

Contents

AUTHOR AFFILIATION AND CONTACT INFORMATION

Corresponding Author:

Daniel Mason-D'Croz (daniel.masondcroz@csiro.au), Commonwealth Science and Industrial Research Organisation (CSIRO). Address: 306 Carmody Road, St Lucia QLD 4067, Australia.

Tel: +61 436.692.889

Sherman Robinson (s.robinson@cgiar.org), International Food Policy Research Institute. Washington, DC

Shahnila Dunston (s.dunston@cgiar.org), International Food Policy Research Institute. Washington, DC

Timothy B. Sulser (t.sulser@cgiar.org), International Food Policy Research Institute. Washington, DC.

ABSTRACT

Simulation models are powerful tools that help us understand, analyze, and explain dynamic, complex systems. They provide empirical methodologies to explore how systems and agents behave and consider how they may change when responding to shocks and stresses. The power of these tools, however, depends on the quality of the data on which they are built. Many complex systems studied in the social sciences, including economic systems, are characterized by sparseness of available data on behavioral characteristics and system outcomes. Generally, there is no single data source that can provide all the necessary information and detail for building a complex, structural, simulation model. Even where good data are available, few datasets are "model ready" without a lot of processing and cleaning. To populate models with data requires significant effort to stitch together a complete, coherent, and model-consistent dataset from a multitude of sources that vary in scope, time-scale, completeness, and quality. Due to information scarcity and variable quality, this challenge is well-suited to a Bayesian approach to efficiently use all available data. To this end, we present a data management system where we apply information theoretic, cross-entropy estimation methods to various FAO agricultural datasets to generate a complete global database of agricultural production, demand, and trade for use in IFPRI's IMPACT model, a global agricultural partial equilibrium multi-market model. We will describe the information theory that serves as the foundation of this methodology, as well as the practical implementation for use in IMPACT.

This data estimation methodology was developed for a partial equilibrium modeling framework, but the principals presented, are applicable to other data processing problems, where there is sparse and poor-quality data (e.g., data for computable general equilibrium models).

Key words: Cross-entropy estimation, Model data management system, Economic Modeling, Agricultural Economics

1 Introduction

Models are potentially powerful tools to help understand, analyze, and explain dynamic systems. They provide systematic methodologies to test how these systems behave and how they may change over time when responding to shocks and stresses. IMPACT is such a tool. It is a partial equilibrium multi-market economic model focusing on global agriculture and food security. It has been used to analyze a variety of questions about potential future challenges to the agriculture and food system including climate change, resource scarcity, technology development, population growth, and economic growth and development. For further details on IMPACT, model design, types of analysis it has been used in, and history and development of the model please see Robinson et al. (2015).

To analyze such complex questions, IMPACT has been greatly disaggregated to provide many coupling points to incorporate data and knowledge from a variety of disciplines (i.e. agronomy, economics, climate science, crop modeling, etc.) to better simulate the complex dynamics of the global agriculture sector. This, however, presents unique data challenges. Data sources vary in scope, time scale, completeness, and quality. Thus, significant effort is required to manage this data to ensure consistency within the model across all these data sources. Due to the scale of IMPACT (158 countries, 62 commodity markets, and long-term time horizons) and the need to update the model on a regular basis to ensure policy relevance, it was critical to develop a systematic and efficient methodology to manage the IMPACT database.

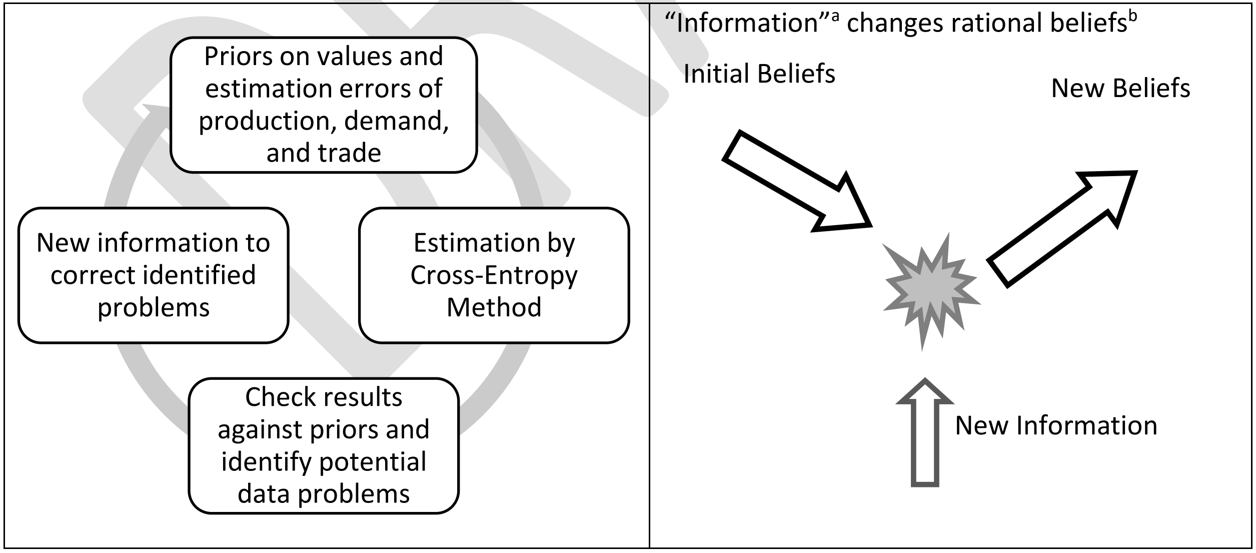
We will explain an efficient data management system that has been applied for IMPACT. In so doing, we will first summarize the theoretical underpinnings of information science upon which our approach is based. We will then describe the data problem and data requirements of IMPACT, and then describe the process used to combine and synthesize all of the disparate data sources for building IMPACT’s base year dataset. The data estimation problem we present will be focused on IMPACT, a partial equilibrium model. Nevertheless, the challenges of building consistent and harmonized datasets is a general modeling challenge, and similar methodologies can and have been applied in other modeling environments (Robinson et al., 2001; Arndt et al., 2002; Go et al., 2016).

2 METHODS AND DATA

2.1 REVIEW OF THEORY AND LITERATURE

The IMPACT data management system follows a Bayesian work plan, where new information can be efficiently added by adjusting appropriate priors and error estimates (Figure 1). Ultimately the goal of this process is to recover parameters and data that we have observed imperfectly. This process has a goal of estimating parameters, as opposed to predicting them, in contrast to the standard statistical approach where there is more data available. It is a powerful methodology that systematically identifies incomplete or unlikely priors, by testing them with all available information while making few assumptions on information we do not have. When this information suggests our initial priors are incorrect, we adjust them based on this information. It can also highlight where additional information is likely needed when certain priors become more unlikely.

Figure 1 Illustrating the Bayesian Work Plan


Priors on values and estimation errors of production, demand, and trade
Estimation by Cross-Entropy Method
Check results against priors and identify potential data problems
New information to correct identified problems
"Information"[a] changes rational beliefs[b]
Initial Beliefs
New Beliefs
New Information

Notes: [a] "Information" is whatever leads to a change in "beliefs".
[b] "Rational" means agents rely on "information" or "evidence" in making decisions.
Source: Adapted from a presentation given by Ariel Caticha (2010)

This methodology is based on information theory and is powerful in large part because it is flexible, allowing us to use different types of information (extremely valuable when working in a data sparse environment). However, to do this it requires us to consider how we measure the content of any new piece of information. The informational content in information theory is determined by how much new information is added to our understanding of the data. If we have high confidence in a prior, and new information confirms our prior, then the added content is relatively low. In contrast, if the information forces us to reexamine our prior, then the content of this additional information is high, as it is changing our current understanding of the data. Claude Shannon (1948), while working at AT&T, developed this measure of "information content", which can be summarized by the following equations, where h is the content of information, and p is the probability that our prior is correct. If the prior is certain (p=1) then the content of any additional information is 0, and if the prior is certain to be wrong (p=0), the content of this new information is infinite.

$$h(p) = \log(1/p)$$
$$IF\ p = 1, then\ h(p) = 0$$
$$IF\ p = 0, then\ h(p) = \infty$$

This can be further developed into an entropy measure, which allows us to estimate the expected information content for a series of information events (k), before they arrive in the following way.

$$H(p) = \sum_{k=1}^{n} p_k \cdot h(p_k) = \sum_{k=1}^{n} p_k \cdot \log(p_k)\,, and$$
$$\sum_{k} p_k = 1$$

Following E.T. Jaynes (1957, 1982, 2003), this entropy measure was applied in the maximum entropy approach, where the idea of estimating probabilities or frequencies was first attempted. Maximum entropy (ME) estimation is achieved by finding the solution from all of the probability distributions, consistent with the estimation constraints, which maximizes Shannon's entropy metric (maximum uncertainty). We then apply Kullback-Leibler's cross entropy (CE) approach, where the estimation problem was redefined as estimating "divergence"[1] of estimated probabilities, which satisfy various constraints or conditions, from the original prior.

$$\begin{gathered}
Minimize: \\
\sum_{k}(p_k \cdot \log(p_k/\bar{p}_k)) = \sum_{k} p_k \cdot (\log(p_k) - \log(\bar{p}_k)) , \\
where\ \bar{p}\ is\ the\ prior\ probability
\end{gathered}$$

Both ME and CE are similar approaches, and in fact, when the prior is specified as a uniform distribution, the CE estimate is equivalent to the ME estimate. This is a particularly useful relationship when working in a sparse data setting, where often the most appropriate assumption is the uniform distribution, which essentially asserts that all "events" are equally unlikely to occur.

To estimate using cross entropy it is necessary to define the type of information that can be directly used. Two types of information are needed: (1) prior distributions of the probabilities of events, and (2) moments within these distributions. The second type of information can include a wide variety of data types, and can be specified as inequalities, data points with errors, or summary statistics like means, medians, or quantiles. Once this information is collected, the CE estimation is done in the following way:

$$\begin{gathered}
Minimize\ p_k: \\
\sum_{k=1}^{K}\left(p_k \cdot \ln\left({}^{p_k}/_{\bar{p}_k}\right)\right) \\
subject\ to\ constraints\ (information)\ about\ moments \\
\sum_{k=1}^{K}\left(p_k \cdot x_{t,k}\right) = y_t \\
and\ the\ adding\ up\ constraint\ (finite\ distribution) \\
\sum_{k=1}^{K}(p_k) = 1
\end{gathered}$$

To estimate the point of maximum entropy in our relationship we need to use Lagrange multipliers (L), which we calculate with the following equation.

[1] Divergence is not a measure of distance., and is not symmetric and thus does not satisfy the triangle inequality

$$L = \sum_{k=1}^{K}\left(p_k \cdot \ln\left({}^{p_k}/_{\bar{p}_k}\right)\right) + \sum_{t=1}^{T}\left(\lambda_t \cdot \left(y_t - \sum_{k=1}^{K}\left(p_k \cdot x_{t,k}\right)\right)\right) + \mu\left(1 - \sum_{k=1}^{K}(p_k)\right)$$

The first order conditions of the CE estimation are therefore:

$$0 = \ln p_k - \ln \bar{p}_k + 1 - \sum_{t=1}^{T}\left(\lambda_t \cdot \mathrm{x}_{\mathrm{t,k}}\right) - \mu$$

$$0 = y_t - \sum_{k=1}^{K}\left(p_k \cdot x_{t,k}\right)$$

$$0 = 1 - \sum_{k=1}^{K}(p_k)$$

This leads to the following equation, which can be thought of a non-parametric Bayesian estimator, transforming the prior and sample information into posterior estimates of probabilities. Ω is called a "partition function" and normalizes the estimated probabilities so that they sum to one. If all the constraints are nonbinding, the lambdas will equal zero, and the estimated posterior ($\tilde{p}_k$) are equal to the prior ($\overline{p_k}$). When this happens then the estimation procedure has added no additional information. If however, the constraints are binding, then the estimated weights will depend on the prior, the value of the lambdas, and the values of the data (X).

$$\tilde{p}_k = \frac{\bar{p}_k}{\Omega(\lambda_1, \lambda_2, \ldots, \lambda_T)} \cdot e^{\sum_{t=1}^{T}(\tilde{\lambda}_t \cdot x_{t,k})}$$

$$where$$

$$\Omega(\tilde{\lambda}) = \sum_{k=1}^{K}\left(\bar{p}_k \cdot e^{\sum_{t=1}^{T}(\tilde{\lambda}_t \cdot x_{t,k})}\right)$$

This gives us a method to estimate probabilities from information; however, in economics, and specifically for IMPACT, we want to estimate parameters to build a balanced and consistent data set. To move from estimating probabilities to estimating parameter values we must adjust how the errors are specified and convert our problem of estimating errors to one of estimating probabilities. First, we need to specify the information we have available, such as parameters (e.g. areas, production, demand, trade, etc.), technology coefficients (e.g. yields, input-output coefficients, etc.), and a prior distribution of the measurement error the means, standard errors, and whether we can assume an informative or uninformative prior distribution. We generate our initial prior as a best estimate of all available data be it values or technology coefficients, using a combination of historical statistics and expert knowledge. Next we need to define the error in our estimation. The error can be assigned to either the technology coefficients or the values, depending on what data is available. The errors can be specified as either additive or multiplicative. For IMPACT's data estimation, we specified additive errors, which allows the error to be positive or negative, which can potentially change the sign of the estimated value, a useful characteristic in the data we are currently using where numbers can change from positive to negative (e.g. net trade). The following set of equations explain the generic error specification of an additive error, where x is the estimated value, $\bar{x}$ is the prior, e is the error, $\bar{v}$ is the error support set, and W are the probabilities that will be estimated.

$$x_i = \bar{x}_i + e_i$$

$$e_i = \sum_k \left(W_{i,k} \cdot \bar{v}_{i,k}\right), where\ 0 \leq W_{i,k} \leq 1\ and\ \sum_k W_{i,k} = 1$$

The support set ($\bar{v}$) gives us the technique to move from estimating values in natural units (i.e. hectares, tones, etc.) to the information approach where the parameters are estimated as probabilities. The support set is defined based on the available knowledge of the prior distribution, and can range from uninformative to varying levels of informative priors (depending on knowledge of the error distribution). Table 1 summarizes what information is needed and how to specify the support set and priors.

Table 1 Specifying the Support Set

Information Needed	Priors	Support Set
Uninformative Prior		
Approximate the uniform distribution applying bounds on the error at (±3s[a])	$Variance: \sigma_i^2 = \sum_k \left(\bar{w}_{i,k} \cdot \bar{v}_{i,k}^2\right), \quad where\ \bar{w}_k = \frac{1}{7}$ $\sigma_i^2 = \frac{s^2}{7} \cdot (0 + 4 + 1 + 1 + 4 + 9) = 4s^2$	$\bar{v}_{i,1} = -3s$ $\bar{v}_{i,2} = -2s$ $\bar{v}_{i,3} = -s$ $\bar{v}_{i,4} = 0$ $\bar{v}_{i,5} = +s$ $\bar{v}_{i,6} = +2s$ $\bar{v}_{i,7} = +3s$
Informative Priors		
Knowledge of the mean and the variance of the error distribution (2 parameters)	$Mean: \sum_k \left(W_{i,k} \cdot \bar{v}_{i,k}\right) = 0$ $Variance: \sum_k \left(W_{i,k} \cdot \bar{v}_{i,k}^2\right) = \sigma_i^2$ $\sigma_i^2 = \left(\bar{W}_{i,1} \cdot 9\sigma_i^2\right) + \left(\bar{W}_{i,2} \cdot 0\right) + \left(\bar{W}_{i,3} \cdot 9\sigma_i^2\right),$ $where\ \bar{W}_{i,1} = \bar{W}_{i,3} = \frac{1}{18}\ ;\ \bar{W}_{i,2} = \frac{16}{18}$	$\bar{v}_{i,1} = -3\sigma_i$ $\bar{v}_{i,2} = 0$ $\bar{v}_{i,3} = +3\sigma_i$
Knowledge of the mean, variance, skewness, and kurtosis (4 parameters)	$Mean\ and\ Variance\ same\ as\ above$ $Skewness: \sum_k \left(W_{i,k} \cdot \bar{v}_{i,k}^3\right) = 0$ $Kurtosis: \sum_k \left(W_{i,k} \cdot \bar{v}_{i,k}^4\right) = 3\sigma_i^4$ $\sigma_i^2 = \left(\bar{W}_{i,1} \cdot 9\sigma_i^2\right) + \left(\bar{W}_{i,2} \cdot \frac{9}{4}\sigma_i^2\right) + \left(\bar{W}_{i,3} \cdot 0\right) + \left(\bar{W}_{i,4} \cdot \frac{9}{4}\sigma_i^2\right) + \left(\bar{W}_{i,5} \cdot 9\sigma_i^2\right)$ $3\sigma_i^4 = \left(\bar{W}_{i,1} \cdot 81\sigma_i^4\right) + \left(\bar{W}_{i,2} \cdot \frac{81}{16}\sigma_i^4\right) + \left(\bar{W}_{i,3} \cdot 0\right) + \left(\bar{W}_{i,4} \cdot \frac{81}{16}\sigma_i^4\right) + \left(\bar{W}_{i,5} \cdot 81\sigma_i^4\right), \quad where$ $\bar{W}_{i,1} = \bar{W}_{i,5} = \frac{1}{162};\ \bar{W}_{i,2} = \bar{W}_{i,4} = \frac{16}{81};\ and\ \bar{W}_{i,3} = \frac{48}{81}$	$\bar{v}_{i,1} = -3\sigma_i$ $\bar{v}_{i,2} = -1.5\sigma_i$ $\bar{v}_{i,3} = 0$ $\bar{v}_{i,4} = +1.5\sigma_i$ $\bar{v}_{i,5} = +3\sigma_i$

Notes: [a] s is a constant and is used to approximate the uniform distribution

2.2 DATA

IMPACT is a large and highly disaggregated global economic model simulating more than 60 commodity markets in 158 countries. Additionally, this economic model is coupled with water models to estimate the effects of water availability on agricultural productivity. This coupling requires further disaggregating production by irrigated and rainfed production systems, as well as disaggregating production into 320 sub-national geographical units, called Food Production Units (FPUs), which are defined as the intersection of country boundaries and watersheds (Robinson et al., 2015). Table 2 summarizes IMPACT's initial data requirements.

Table 2 IMPACT data requirements in the model base year

Data Source	Geographic Scope	IMPACT Parameter	Commodity Requirement	Unit
OECD-AMAD[a]	Global	World Prices	All commodities	USD/metric tonnes (mt)
WDI[b] and CIA World Factbook[c]	National	Population	-	Million
		GDP	-	Billion USD, PPP
FAOSTAT[d] Commodity Balances	National	Total Supply	All commodities	000 mt
		- Animal Numbers	Livestock only	000 producing animals
		- Harvest Area	Crops only	000 hectares (ha)
		- Yield	Crops & livestock	mt/ha
		Total Demand	All commodities	000 mt
		- Food Demand	All commodities	000 mt
		- Feed Demand	All commodities	000 mt
		- Intermediate Demand	All commodities	000 mt
		- Other Demand	All commodities	000 mt
		Stock Change	All commodities	000 mt
		Net Trade	All commodities	000 mt
FAOSTAT Food Supply	National	Calorie Availability	-	kcal/person/day
		Food Supply Quantity	Food commodities	kg/capita/yr
		Food Supply	Food commodities	kcal/commodity/person/day
FAO AquaStat[e] and OECD[f]	National	Total Irrigated Area		000 ha
		Irrigated Crop Area	Crops only	000 ha
IFPRI SPAM[g]	FPU (aggregated from pixels)	By production system (irr/rfd):		
		- Harvest Area	Crops only	000 ha
		- Yield	Crops only	mt/ha
		- Production	Crops only	000 mt

Notes: [a] OECD's Agricultural Market Access Database (OECD, 2010)
[b] World Bank's World Development Indicators (World Bank, 2014)
[c] U.S. CIA World Factbook used when data missing from WB (US CIA, 2014)
[d] FAO's FAOSTAT Database (FAO, 2015a)
[e] FAO's AquaStat Database (FAO, 2015b)
[f] OECD Agriculture Statistics (OECD 2014)
[g] IFPRI's Spatial Production Allocation Model (You et al., 2014)

The challenge of merging all this data is significant. Each dataset has its own metadata, with varying geographic and commodity focus and definitions. This requires developing protocols to stitch together all this data to the regional and commodities used in IMPACT. Additionally, IMPACT provides a logical framework, which serves as additional data estimation constraints, as the product of this data processing exercise should be the base year solution of the model. The following rules of IMPACT's equilibrium solution are applied as data estimation constraints:

1. There is an equilibrium for every commodity market, which is defined in IMPACT as:
    a. $\sum_{cty} TotalDemand = \sum_{cty} TotalSupply$, and
    b. $\sum_{cty} NetTrade = 0$
2. Nationally, there must be a perfect accounting of production, demand and trade:
    a. $NetTrade = TotalSupply - TotalDemand$
3. Production is defined by 2 general accounting rules
    a. For crops and livestock: $Supply = Area\ (or\ animals) \times Yield$
    b. For processed commodities, the mass of inputs must be equal to or greater than the mass of outputs (including waste): $\sum_{input} IntermediateDemand \geq \sum_{output} Supply$
4. Total demand is the sum of all demand types
    a. $TotalDemand = Food + Feed + Intermediate + Other$

However, many of these conditions are not likely to be found in the dataset. Additionally, the various geographic and commodity definitions across data sets poses a major challenge in fitting the data to IMPACT's logical framework, as does the poor quality and completeness of the data available. IMPACT works at a more spatially disaggregated level than at the country-level at which most of the statistics used in IMPACT are reported. This requires then not only balancing the national statistics using the above rules, but also then disaggregating production and demand. The data that is used to disaggregate the national numbers, come from different datasets (FAO's AquaStat, and IFPRI's SPAM), which are not assured to match up with our national statistics (FAOSTAT). To work around this problem the data from AquaStat and SPAM are treated as shares, which are then used to calculate disaggregated numbers, such that the following is true:

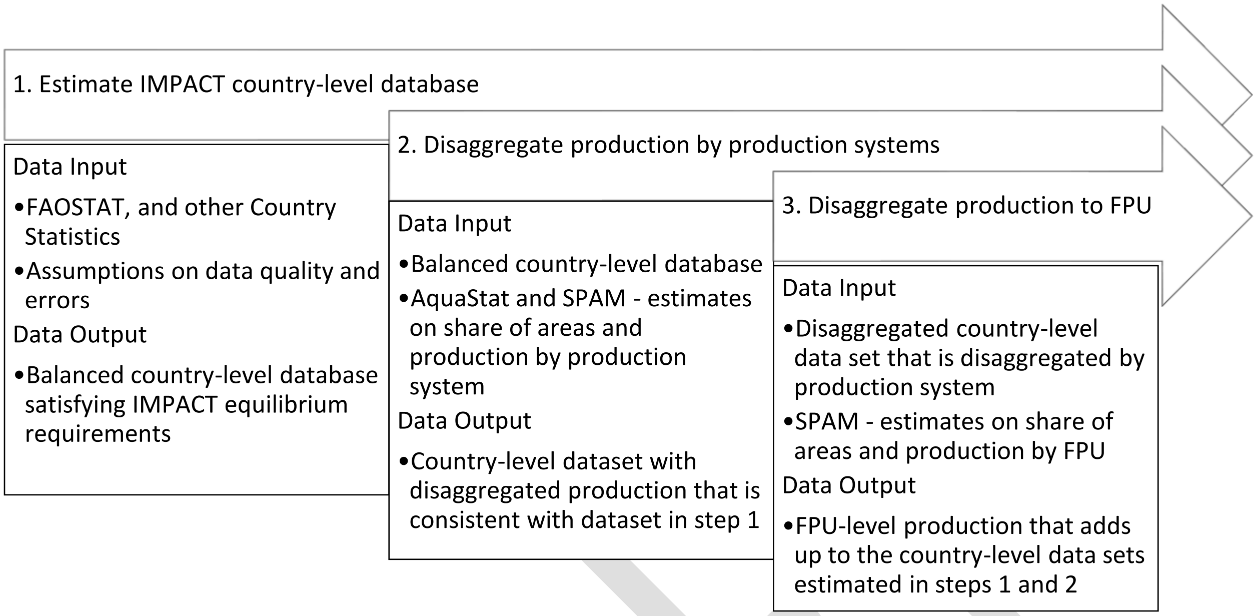
5. $CountrySupply = \sum_{landtype}\left(\sum_{fpu}(Supply)\right)$.

In a smaller model, one could imagine doing all this balancing and cleaning process carefully by hand or given sufficient time in an iterative adjustment process that slowly converges to a solution. Nevertheless, due to the size and complexity of the database needed for a global economic model like IMPACT and the desire to be able to semi-regularly update the database, it was necessary to implement an information efficient data management system, which we will describe in the following section.

2.3 ESTIMATING A CONSISTENT BASE YEAR DATASET FOR IMPACT

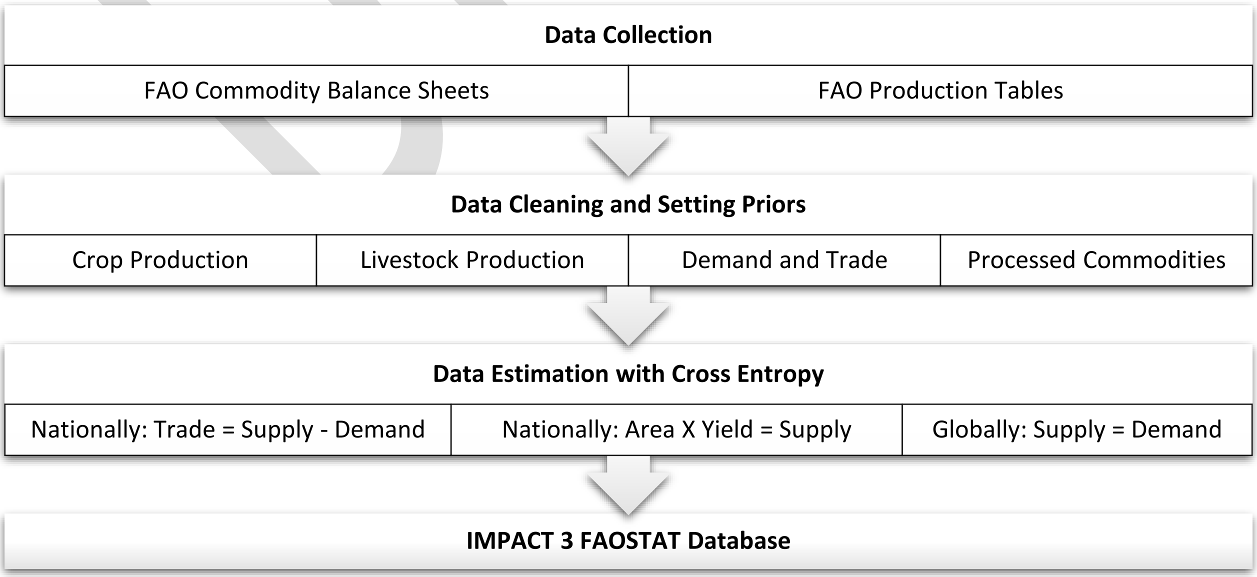
In the previous sections, we summarized the information theory upon which our data estimation is based, as well as the range of data inputs that are currently being used. This section will focus on explaining the practical steps that are required to implement the theory in an estimation program written in GAMS. The IMPACT data management applies the cross-entropy approach in 3 stages that progressively harmonizes country-level data and subsequently disaggregates it sub-nationally to incorporate data on crop production systems (irrigation and rainfed) and watersheds. In this process the results of the solution of previous stages of estimation serve as priors and constraints to subsequent stages of estimation (Figure 2). Breaking up the overall estimation process into 3 smaller estimation problems has the benefit of allowing each problem to be designed modularly, which allows each step to be run in isolation, or combined if desired. For example, if one wanted to run a partial equilibrium model without IMPACT's sub-national units it would be possible to run the 1st stage of estimation, without running the 2nd and 3rd stages.

Figure 2 The 3 stages of estimating a complete IMPACT ready data set


1. Estimate IMPACT country-level database
Data Input
•FAOSTAT, and other Country Statistics
•Assumptions on data quality and errors
Data Output
•Balanced country-level database satisfying IMPACT equilibrium requirements
2. Disaggregate production by production systems
Data Input
•Balanced country-level database
•AquaStat and SPAM - estimates on share of areas and production by production system
Data Output
•Country-level dataset with disaggregated production that is consistent with dataset in step 1
3. Disaggregate production to FPU
Data Input
•Disaggregated country-level data set that is disaggregated by production system
•SPAM - estimates on share of areas and production by FPU
Data Output
•FPU-level production that adds up to the country-level data sets estimated in steps 1 and 2

The first stage, where we reconcile various country-level data sources into an IMPACT consistent dataset, is the largest of the 3 estimation problems, requiring most of the data collection and data cleaning, and for this reason, in this paper we will focus on this stage of the data estimation. Figure 3 summarizes the different steps involved in the first stage of estimation, which involves significant data cleaning, setting priors, and finally the cross-entropy estimation.

Figure 3 First data estimation problem: Balancing FAOSTAT


Data Collection
FAO Commodity Balance Sheets
FAO Production Tables
Data Cleaning and Setting Priors
Crop Production
Livestock Production
Demand and Trade
Processed Commodities
Data Estimation with Cross Entropy
Nationally: Trade = Supply - Demand
Nationally: Area X Yield = Supply
Globally: Supply = Demand
IMPACT 3 FAOSTAT Database

Most of the data used to set our initial priors is drawn from various FAO datasets (i.e. FAOSTAT's ProdStat and Commodity Balance Sheets). The data for the years 2004-2006 was downloaded and loaded in GAMS and mapped from FAO to IMPACT regions and commodities. We did this to allow us to better identify outliers in the dataset, and to capture data that might have been unreported in any year. Once mapped, we calculate a 3-year average centered on 2005, IMPACT's base year, and using a suite of diagnostics statistics we started the process of recapturing missing data. For example, if we had a value for area and production, but the value for yield was missing, we estimated the yield by dividing the production by the area. Where the challenge of recapturing missing or replacing erroneous data was more difficult, we used alternative data sources (e.g. national statistics) or estimated priors based on global or regional averages or medians. For example, we replaced extreme low values for crop yields by estimating a minimum yield floor that is drawn from the global distribution of yields. This preliminary stage of data cleaning is essential, in that we are incorporating as much information as possible in our initial priors. The better the priors the better our data estimation will be. Poorly informed priors, on the other hand, may lead to an infeasible estimation problem, where it is not possible to find a solution that satisfies all the constraints. From our experience, it is almost impossible to get a solution on the first attempt when working with large data sets. Nevertheless, the infeasible solution outputs from GAMS solvers[2] combined with good diagnostic code can point to which priors are especially problematic and need to be reviewed (Bayesian Workplan).

After we cleaned the data inputs, and established priors on parameter values, we needed to determine if there is any information on parameter error distributions. To do this we decided on a applied a tiered hierarchy of data quality for the data inputs used. This hierarchy describes, the quality of the data,

[2] We used various solvers including CONOPT, IPOTH, KINITRO, and MOSEK while testing the estimation model. Ultimately we preferred using CONOPT, and MOSEK with respect to speed and the usefulness of solver error outputs.

and in which parameter values we have higher confidence. This hierarchy was developed through extensive conversation with commodity experts, and with the statistics division at FAO. The tiers are as follows: high confidence (areas[3], and supply), medium confidence (food demand, livestock feed demand, and intermediate demand), and low confidence (other demand, stock change, exports, and imports). For the high and medium confidence tiers we applied a 5 support element set, with the difference between the high and medium based on the allowable size of the error. The low confidence tiers were allowed even larger errors and were assumed to have uninformative error distributions. Table 3 summarizes the estimated parameters, along with assumptions on the error distribution. Using the σ specified below for each parameter we calculate $\bar{w}$ (prior on error probabilities), and $\bar{v}$ (support set for error) for each parameter following the equations explained in Table 1.

Table 3 Specifying cross-entropy error estimation

Parameter	Assumption on Error Distribution	Element Support Set	Prior on σ
Area (ARA)			$0 \cdot ARA_{j,cty}$
Supply (QS)	Informative	5[a]	$\pm 0.1 \cdot QS_{j,cty}$[b]
Food Demand (QF)	Informative	5	$\pm 0.5 \cdot QF_{j,cty}$
Livestock Feed Demand (QL)	Informative	5	$\pm 0.5 \cdot QL_{j,cty}$
Intermediate Demand (QINT)	Informative	5	$\pm 0.5 \cdot QINT_{j,cty}$
Other Demand (QOTH)	Uninformative	7	$\pm 0.5 \cdot QOTH_{j,cty}$
Stock Change (QST)	Uninformative	7[c]	$\pm \max\left[0.05 \cdot \left(QS_{j,cty} + QD_{j,cty}\right), \left\lvert QST_{j,cty}\right\rvert\right]$
Imports (QM)	Uninformative	7[d]	$\pm \max\left[0.5 \cdot \left(QM_{j,cty}\right), 0.5 \cdot \left(QD_{j,cty}\right)\right]$
Exports (QE)	Uninformative	7	$\pm \max\left[0.5 \cdot \left(QE_{j,cty}\right), 0.5 \cdot \left(QS_{j,cty}\right)\right]$

Notes: j stands for crop, and cty stands for country

[a] All of the informative priors use the 2 parameter informative prior

[b] Palm Oil Fruit is the lone exception with an allowable deviation of 0.15

[c] Stock changes have a large potential deviation as this data type has very low data quality

[d] Imports and exports need a larger deviation and are based on either the base trade flows or demand (or supply)

[3] Currently, we assume FAO harvest area as a binding constraint, and don't include this parameter in the cross-entropy estimation.

Once the priors and error distributions have been specified all that remains is to run the cross-entropy estimation model, which is defined in Box 1. The model can be broken up into 3 main pieces:

1. We define the equations focused on the error terms.
2. We define the equations that estimate parameter values in natural units (i.e. tons)
3. We define equations which specify the estimation constraints

Box 1 Country-level estimation model equations

Solve Minimizing CNTRPY

Entropy Equation

$$CNTRPY == \sum\left[W_{j,cty,k} \cdot \left(\log\left(W_{j,cty,k} + \Delta\right) - \log\left(\overline{w}_{j,cty,k} + \Delta\right)\right)\right], where\ \Delta = 1e^{-6}$$

Sum of Errors by crop (j) and country (cty)

$$\sum_{k} W_{j,cty,k} == 1$$

Error Equation

$$ERR_{j,cty} == \sum\left(W_{j,cty,k} \cdot \bar{v}_{j,cty,k}\right)$$

Supply Equation

$$QS_{j,cty} == \overline{qs}_{j,cty} + ERR_{j,cty}$$

Demand Equations

$$QF_{j,cty} == \overline{qf}_{j,cty} + ERR_{j,cty}$$
$$QL_{j,cty} == \overline{ql}_{j,cty} + ERR_{j,cty}$$
$$QINT_{j,cty} == \overline{qint}_{j,cty} + ERR_{j,cty}$$
$$QOTH_{j,cty} == \overline{qoth}_{j,cty} + ERR_{j,cty}$$
$$QD_{j,cty} == QF_{j,cty} + QL_{j,cty} + QINT_{j,cty} + QOTH_{j,cty}$$

Trade Equations

$$QST_{j,cty} == \overline{qst} + ERR_{j,cty}$$
$$QE_{j,cty} == \overline{qm} + ERR_{j,cty}$$
$$QM_{j,cty} == \overline{qe} + ERR_{j,cty} Net\ Trade\ Equation:$$
$$QN_{j,cty} == QE_{j,cty} - QM_{j,cty}$$

Country Supply Demand Balance

$$QS_{j,cty} == QN_{j,cty} + QD_{j,cty} + QST_{j,cty}$$

Constrain Exports

$$QS_{j,cty} \geq QE_{j,cty}$$

Global Net Trade Balance

$World\ Production\ must\ equal\ World\ Demand$:

$$\sum_{cty} QN_{j,cty} = 0$$

Yield Equation

$$Yld_{j,cty} = \frac{QS_{j,cty}}{ARA_{j,cty}}, where\ Yld_{j,cty} \geq Yield\ Floor$$

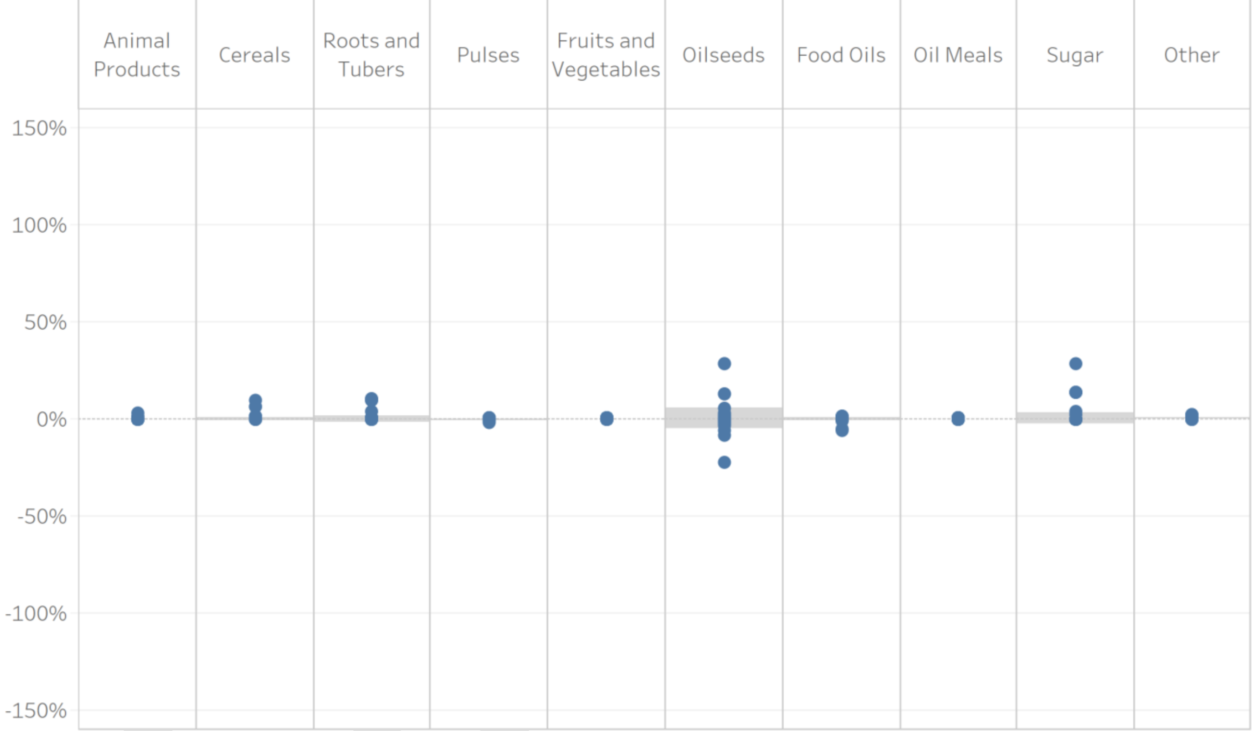
It is possible to run all commodities simultaneously. However, as each commodity is defined as an independent estimation problem (i.e. the adding up constraints are crop specific, maize production is not a function of vegetable production). This allows each commodity to be solved independently, which allows for easier data checking and error spotting.

3 RESULTS

The cross-entropy estimation of a model consistent dataset is not the final step of our Bayesian Workplan. It produces a proto-database, but before we use it in IMPACT we need to review the results and determine if the results of the estimation deviate from our priors in an understandable and acceptable fashion. Did the estimation confirm our assumptions of the data we worked with, or did they raise new questions that required us to find and introduce new information to improve our priors and thereby the proto-database produced by the next iteration of cross-entropy estimation. In this section we will review examples where the estimation process provided low information, meaning the results did not diverge radically from the priors, as well as when it provided high information along with subsequent iterations after having applied additional information to our priors.

Figure 4 summarizes the deviations from the prior for commodity supply after one round of data estimation, which we will call **R1** to distinguish it from subsequent solutions. We should note that **R1** was not the first iteration of our Bayesian Workplan. It is one of the later iterations selected because it allowed us to easily isolate a single data issue to illustrate how the estimation program helped to identify and resolve data errors. In general in **R1**, the estimation process did a good job, with most of the estimations in aggregate not adding a lot of additional information, with most results deviating by less than 10 percent from our priors. A reasonably successful estimation given that we were solving for 5,388 different supply estimates across all commodities and countries. The mean deviation across all commodity groups was close to 0, with the largest observed standard deviation for oilseeds at ±5.35.

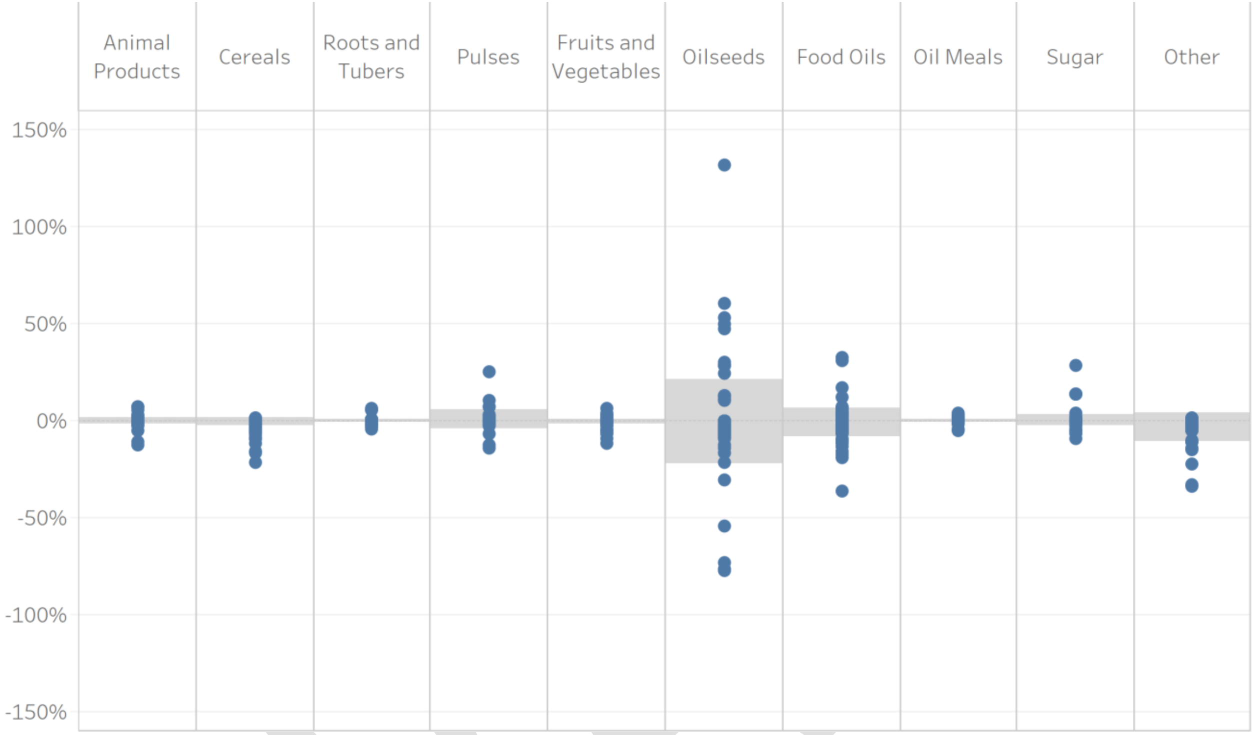
Figure 4 Commodity supply percent deviation from Prior by commodity group



Note: Blue dots represent country supply for commodities within each commodity group (e.g. U.S. wheat in cereals)
Grey boxes represent area within ± σ²

The relatively small deviation for supply observed in Figure 4 is in some part by design (see Table 3), in that we had a fairly tight constraint on the allowable error around commodity supply, suggesting that we must also review the results of deviations to other parameters to assess the quality of the proto-dataset from **R1**. Figure 5 summarizes the percent deviation for commodity demand for **R1**. The size of deviations for demand are unsurprisingly higher given our assumptions on the allowable error. In some cases, the increase in error is quite large. For example, the grey box representing the range between ±1 standard deviation for oilseeds in Figure 5 is larger than the range of all deviations for oilseeds supply seen in Figure 4, suggesting that the priors for the oilseed commodities need to be revised.

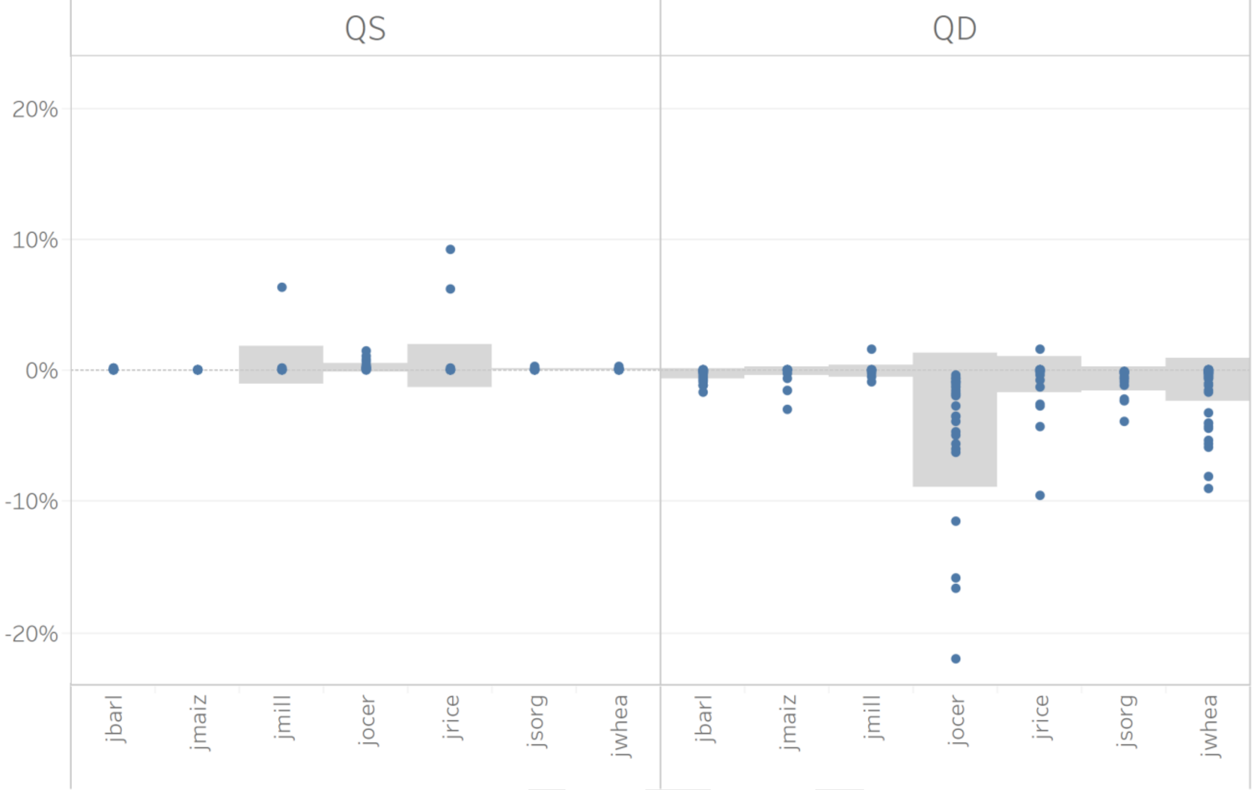
Figure 5 Commodity demand percent deviation from Prior by commodity group



Note: Blue dots represent country demand for commodities within each commodity group (e.g. U.S. wheat in cereals)
Grey boxes represent area within ± σ²

However, for many commodity groups (e.g. animal products, and cereals) the deviation from the priors for demand are generally small, and comparable to those observed for commodity supply. Figure 6 summarizes the deviations for supply and demand for cereals. While deviations from the initial priors for commodity demand are on average larger than those for supply, the deviation across most of the cereal commodities is ±5 from the prior, suggesting that the initial priors are fairly informative, and that the estimation process has added limited additional information. Larger deviations within the cereal grouping can be observed for jocerl (other cereals). Given that this is an aggregate commodity, which encompasses a fairly heterogenous mix of more minor cereals (e.g. oats, rye, and triticale) and pseudocereals (e.g. amaranth, buckwheat, and quinoa), it isn't surprising that the initial prior isn't as good as it is for highly commercial grains like maize, where the initial priors on supply and demand appear to be well informed.

Figure 6 Comparing percent deviation for commodity supply and demand for cereals

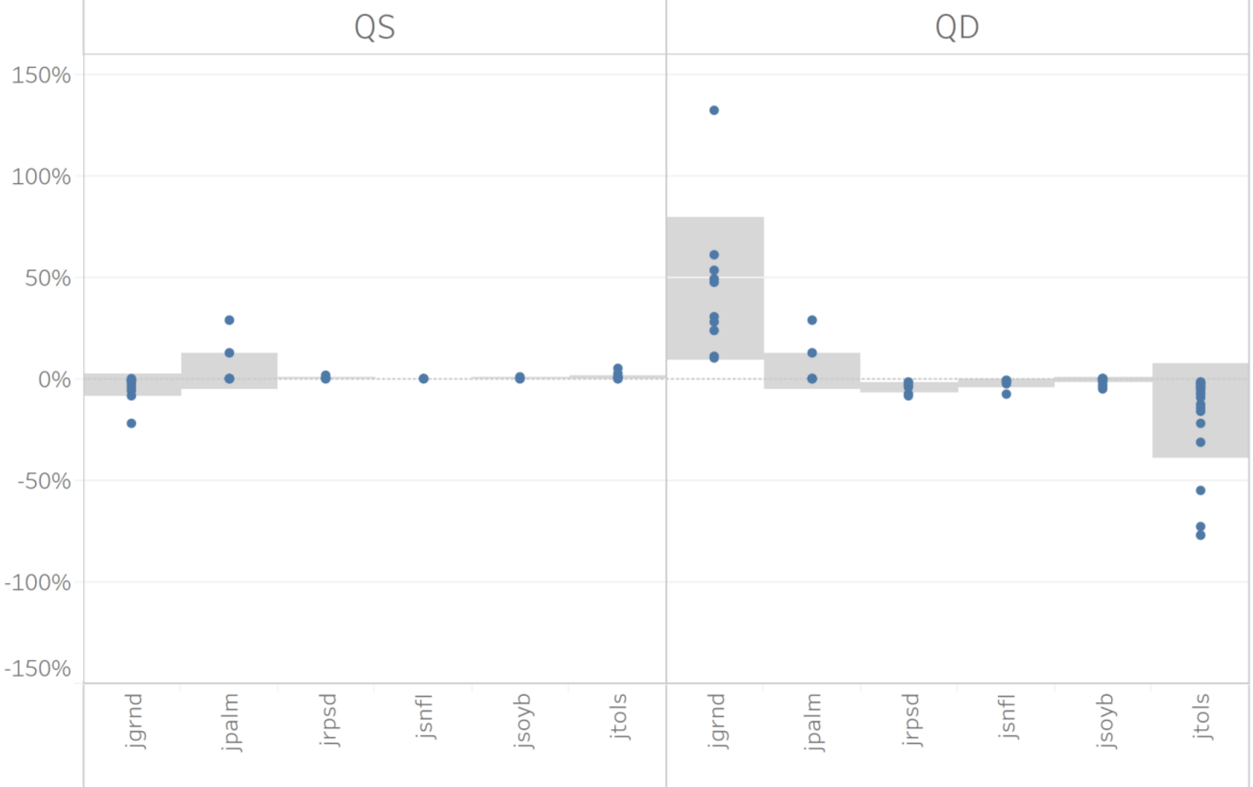


Note: Blue dots represent country supply for commodities within each commodity group (e.g. U.S. wheat in cereals)
Grey boxes represent area within ± σ^2
jbarl = barley; jmaiz = maize; jmill = millet; jocerl= other cereals; jrice = rice; jsorg = sorghum; jwhea = wheat

While the deviations are fairly small within the cereal commodities compared to other commodities like oilseeds, we can still find individual priors within each of the commodities which are not a great fit. When we drill down and explore these outliers we find generally they are countries that are small producers and consumers, such that while the deviation is large in percentage terms it is fairly small in physical units. For example, the largest deviation for cereals (excluding jocerl) in Figure 6 is for rice demand in Saudi Arabia where we see a deviation of about 9.5 percent from our prior. This deviation is larger than would be ideal, but at the global scale that IMPACT operates such a deviation is relatively small, given that Saudi Arabia consumes less than 0.2 percent of global rice demand. Undoubtedly, additional information could be applied to rice markets that would allow for greater fit but given some of the larger deviations in other commodity groups we decided that rice markets were a lower priority in subsequent iterations of our Bayesian Workplan.

Returning to Figure 4 and Figure 5 the commodity group with the largest deviation, suggesting that for this commodity group the cross-entropy estimation solution has high information content, suggesting that we should review and revise our priors. Figure 7 replicates for oilseeds the view comparing supply and demand for cereals in Figure 6. As for cereals we see that the deviations from the prior for demand are larger than for supply. However, unlike for cereals the deviations even for supply in Figure 5 are large compared to all of the other commodity groups. As we drill down within the oilseed category in Figure 7 we can see that much of this can be pinned to three commodities (jgrnd, jpalm, and jtols).

Figure 7 Comparing percent deviation for commodity supply and demand for oilseeds



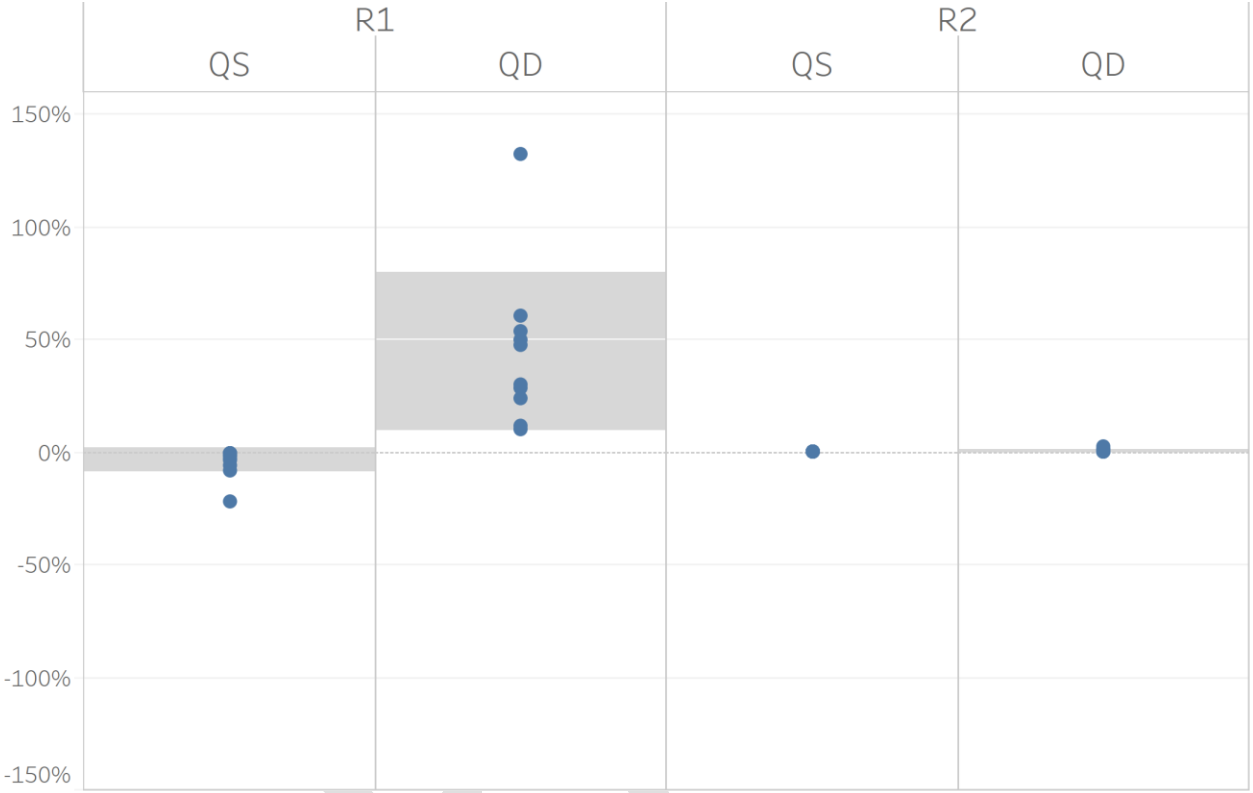
Note: Blue dots represent country supply or demand for commodities within each commodity group
Grey boxes represent area within ± σ²
jgrnd = groundnut; jpalm = palm; jrpsd = rapeseed; jsnfl= sunflower; jsoyb = soybean; jtols = other oilseeds

Two of these three can be explained in similar ways as we did for observed deviations in the cereals commodities. Palm fruit production (jpalm) is treated as a non-traded good, where all production is demanded by domestic palm processors (palm oil is traded in IMPACT), which explains why the deviation for demand is identical to that of supply. Additionally, as palm fruit production is dominated by a small number of countries (Indonesia and Malaysia account for nearly 75 percent of global production), then it can be difficult to make adjustments among many smaller producers without the deviations being large in percent terms. In the case of jpalm the largest outliers are mostly found among West African producing countries like Guinea and Cote d'Ivoire, which account for less than 2 percent of global production. Other oilseed production (jtols) is similar to jocerl, in that it is a heterogeneous mix of different commodities (e.g. coconuts and olives), which when combined unsurprisingly leads to less informative priors. The fact that the deviations for jtols is significantly larger than for jocerl, suggests that even though it is a heterogenous mix there would be value in either improving the priors for this aggregation, perhaps through further disaggregation of the commodity group.

Of the three oilseed commodities, one cannot be discarded with these explanations. Groundnut production (jgrnd) is also the commodity group with the largest deviation. Unlike jtols, jgrnd is a mostly homogenous commodity, and unlike jpalm it is not anywhere near as concentrated with the largest producer (China) accounting for a little more than a third of global production. Additionally, the range of deviations for demand showed that all countries' priors were too low and for the most part too high for supply. This suggested that there was a mismatch between the supply and demand datasets, and when we returned to the original data we found that this was exactly the case. The data for groundnut supply was designated "in shell", whereas the data for demand was "shelled, equivalent". In **R1**, then the cross-entropy estimation was able to find a solution the satisfied all of the data constraints while simultaneously spreading the error in groundnut data matching globally, essentially translating all demand data from shelled to "in shell" in a brute force manner. Obviously this solution was far from ideal, and upon discovering our mistake we adjusted our priors for supply to be in shelled equivalent to match the demand data, and ran the cross-entropy estimation a second time (**R2**).

Figure 8 shows the results for groundnuts for supply and demand after we corrected the data mismatch between the two datasets. In **R2**, we see the priors for groundnuts are well informed, with very small deviations (the largest being of 2.3 percent for China), much more in line with what we would expect for a highly commercialized and traded cash crop like groundnuts.

Figure 8 Comparing results for R1 and R2 for supply and demand of groundnuts



Note: Blue dots represent country supply for commodities within each commodity group (e.g. U.S. wheat in cereals)
Grey boxes represent area within ± σ^2

The correction of this data mismatch not only cleared up the deviations for groundnuts, but went a long way to explaining much of the deviations for the oilseeds categories observed in Figure 4 and Figure 5, reducing the range of deviations for the oilseeds category by more than half.

4 Discussion

Our hope with preparing this paper was to demonstrate an efficient data management system that can take various datasets and merge them for use in a large scale global economic model. We believe that it presents a framework to approach the data cleaning and estimation process, necessary for all complex models, that is powerful and solidly based on information theory. It approaches a data estimation problem where there is sparse data, and issues with data quality that challenge the stitching together of datasets with different regional and commodity definitions, in a systematic fashion, that allowed us to iteratively identify, prioritize, and then drill down on and correct questionable priors.

In this paper we showed an example where we were able to identify and prioritize which priors were more questionable than others. We determined that while priors for cereals were not perfect, they were acceptable in the face of the much bigger problem within oilseeds. Drilling down on oilseed commodities we could explain deviations for palm fruit (heavy concentration of production) and other oilseeds (very heterogenous aggregation) such that these were determined to be of lesser importance, while quickly recognizing that groundnuts needed to be the first priority given that the deviations were so large, and for a crop that should have relatively decent global data. In the end, this error was found to be a relatively simple case of a mismatch between definitions of groundnuts between FAO's production tables and the commodity balance sheets that provided data for demand. Nevertheless, it highlighted how our data management system can quickly point out questionable data, and how approaching the data problem in the manner presented in this paper we can improve the data used, with each new iteration of the Bayesian Workplan identifying progressively smaller data errors.

ACKNOWLEDGEMENTS

The cross entropy-estimation module was completed as a part of the Global Futures and Strategic Foresight Project, with funding provided by the CGIAR Research Program on Policies, Institutions, and Markets (PIM), the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS), and the Bill and Melinda Gates Foundation. Additionally, we would like to acknowledge the Commonwealth Science and Industrial Research Organisation (CSIRO), for giving an allowance to the lead author to complete this work during their transition from organizations.

REFERENCES

Arndt, C., S. Robinson, and F. Tarp. (2002). “Parameter Estimation for a Computable General Equilibrium Model: A Maximum Entropy Approach.” *Economic Modelling,* 19: 375–398.

Caticha, Ariel (2010). Max Entropy Tutorial Session 2 – Entropic Inference. PowerPoint Presentation available at: http://djafari.free.fr/MaxEnt2010/slide/Tutorial2_Caticha.pdf

Go, D., H. Lofgren, F. Ramos, and S. Robinson. (2016). “Estimating Parameters and Structural Change in CGE Models Using a Bayesian Cross-Entropy Estimation Approach.” *Economic Modelling*, 52(2016): 790–811. DOI: 10.1016/j.econmod.2015.10.017

FAO (Food and Agriculture Organization of the United Nations). 2015a. FAOSTAT Database. Available at http://faostat3.fao.org.

————. 2015b. AQUASTAT Database. Available at www.fao.org/nr/water/aquastat/main/index.stm.

Jaynes, E. T. (1957). “Information theory and statistical mechanics”. *Physical review*, 106(4): 620–630.

————. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9): 939-952.

————. (2003) *Probability Theory*. Cambridge: Cambridge University Press

OECD (Organisation for Economic Co-operation and Development). 2010. Agricultural Market Access Data Base. [Accessed 11/1/2013]. http://www.oecd.org/site/amad

————. 2014. *Agricultural Policy Monitoring and Evaluation 2014*: OECD Countries. Paris. DOI: 10.1787/agr_pol-2014-en

Robinson, S., A. Cattaneo, and M. El-Said. (2001) “Updating and Estimating a Social Accounting Matrix Using Cross Entropy Methods.” *Economic Systems Research*, 13(1): 47–64.

Robinson, S., D. Mason-D'Croz, S. Islam, A. Guneau, G. Pitois, T. Zhu, C. Ringler, T. Sulser, A. Palazzo, and M.W. Rosegrant. (2015). *The International Model for Policy Analysis of Agricultural Commodities and Trade (IMPACT) Model Description, 3rd Version.* IFPRI Technical Report. Washington, DC: International Food Policy Research Institute.

Shannon, C.E. (1948). "A mathematical theory of communication". *Bell system technical journal*, 27(3), 379-423.

U.S. CIA (2014). World Factbook. Washington, DC. Available at www.cia.gov/library/publications/the-world-factbook

World Bank. (2014). World Development Indicators. Washington, DC. Available at http://data.worldbank.org/data-catalog/world-development-indicators

You, L., U. Wood-Sichra, S. Fritz, Z. Guo, L. See, and J. Koo. 2014. "Spatial Production Allocation Model (SPAM) 2005 v2.0."