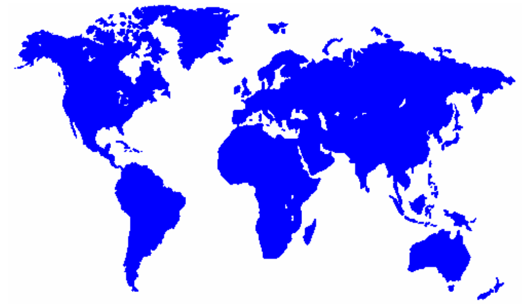


The Data Pre-Processor of International Futures (IFs)

Barry B. Hughes
(with Mohammod T. Irfan on Education)

Version 1.0. Feedback on this living document will be much appreciated.

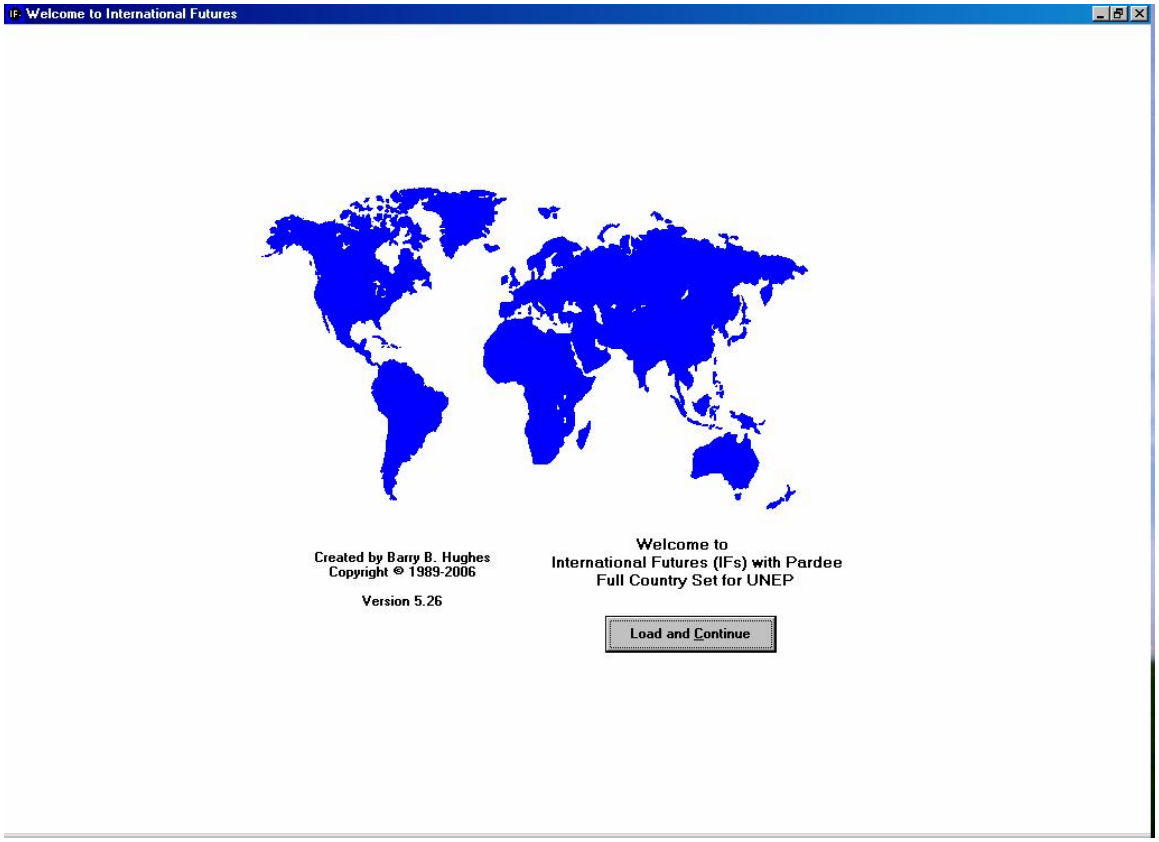
Graduate School
of International Studies
University of Denver
June, 2006

The Data Pre-Processor of International Futures (IFs)

Table of Contents

Abstract

This paper documents the approach taken within the International Futures (IFs) modeling system to initiating and linking the physical and monetary sides of the model. In particular, it provides documentation of the data pre-processor that builds the initial data load for the model.



Created by Barry B. Hughes
Copyright © 1989-2006

Version 5.26

Welcome to
International Futures (IFs) with Pardee
Full Country Set for UNEP

Figure 1. Welcome to IFs.

1. Initializing and Connecting Multiple Modules

The International Futures (IFs) model of long-term, multi-issue global change is imbedded in a largely modeling system that has grown up over time to make the processes of model development and use more efficient and powerful. The portion of that infrastructure known best to most users is the interface, with its data display and analysis, forecast display and analysis, and scenario analysis capabilities.

"Under the hood," however, is another portion of that infrastructure that is fundamentally important to the enterprise. In the IFs system it is known as the data preprocessor. Its primary function and reason for being is preparing initial conditions for the model from data with significant inadequacies.

Preparing an initial data load for a model sometimes requires almost as much work as does creating and maintaining the dynamics of the model. Data inconsistency and data holes require attention; in a model like IFs with physical representations of partial equilibrium sectoral models (agriculture and energy) as well as a general equilibrium multi-sector model represented in value terms, there is the also the need to reconcile the physical and value data.

There are large numbers of missing data that are essential for initializing the base or initial forecasting year on a country by country level. In addition, data are often inconsistent from one source to another or are in different units that must be converted to the dominant ones used in each portion of the model (for example, millions of people, billions of dollars, and billion barrels of oil equivalent). Further, there are often substantial discrepancies between data in the physical models (such as energy or agricultural trade) and those in the economic model (such values of energy, agricultural and total trade). In addition, global totals often do not sum on demand or supply sides (such as trade again).

The preprocessor also has the function of aggregating or disaggregating data as necessary. For many years in the development of IFs, the model was run with countries routinely aggregated in advance into regions of the world – the initial aggregation was into 14 large world countries and groupings of them. In the future it is planned that IFs will be able to process data from sub-regions of countries and either run with representation of such sub-regions or aggregate them into countries.

All large-scale global models and probably most other macro models of global systems have similar problems with missing data and data consistency. The need to periodically change the base year of such models compounds such problems and the regular availability of data revisions and updates can make them a horrendous time sink for modelers. In the GLOBUS world modeling project, for example, projects lasting two or more years or twice undertaken merely to update the database, at the end of which it was already again somewhat out of date.

The IFs preprocessor began essentially as a codification of the rules or algorithms of such data updates and revisions, growing gradually into a complex system of its own that has certainly saved countless person-years of labor and also facilitated regular, near immediate data updates. The pre-processor greatly facilitates both partial data updates as better data become available and rebasing of the entire model to a new initial year (such as the rebasing from 1995 to 2000).

The preprocessor has also facilitated the development of a capability for running the model historically, from 1960 forward. It works with an extensive raw data file for all areas of the model, using data gathered for 1960 through the most recent year available. Such runs from 1960 allow, for instance, the calibration of the model against history and the exploration for and identification of the kinds of events or shocks that might be most likely to confound the model's ability to forecast beyond the present. Such historic runs obviously also require the creation of an initial data load, and because data in 1960 have considerably greater weaknesses than those in 2000 or later, the need for cleaning and processing is even greater.

1.1 General Approaches to Data Problems

Over time a number of approaches have evolved to facilitate recurrent needs in the preprocessor. These include:

- Filling of data holes with values estimated from cross-sectional relationships. What does one do, for instance, with a country that is missing data on variables as essential as total fertility rate or malnourished children? In some cases they can be calculated or estimated from foundational data, as fertility rate might be taken from cohort-specific fertility tables. In other cases, however, there is no obviously related database. In such cases a cross-sectional relationship is generally used in the preprocessor. Because GDP per capita, especially at purchasing power parity, is so often highly correlated with social-system variables, it is by far the most common independent variable in such relationships. Very often it has been found that logarithmic relationships provide the best fit, but IFs uses a variety of forms as appropriate.

- Filling of other data holes with a very small non-zero value. Such a "seed" captures the reality that most values are not absolutely zero even in countries so shown in data, and very seldom in countries with missing data. The seeds, as used in other models such as SARUM (SARU, xxx) also allow the later dynamic forecasts to build on a value in multiplicative as well as additive specifications.

- Balancing global sums, for instance of imports and exports or of inward and outward flows of people. Most often IFs simply averages the two sums and uses the average to normalize the contributions of each country to the respective sums. More generally, normalizations are widely used in the preprocessor.

- Iteration within constraints. Sometimes complex systems of variables, such as those around the input-output matrices of countries and the physical variables that tie to those matrices, interact in ways that do easily allow one data value to be "privileged" with respect to others that are in initial contradiction to them. In such instances, the preprocessor sometimes iterates across the system until various constraints on the variables within it are met.

- Integration of historic flows to calculate an estimate of stock levels for initialization of forecasts. For instance, stocks of migrants and of FDI can be computed in this manner.

- Calculation of initial growth rates for variables from the growth patterns over a period prior to the initial forecast year. Because many growth rates, such as GDP or urban population, vary considerably from year to year, using a period from 2-10 years for initialization is generally preferable to using a single year. In some cases, such as GDP, growth rates are calculated from the values rather than the rates, always using compound annualized calculations over a period rather than the simple rate of change over the entire period.

1.2 Sequence Design of the Preprocessor

Sequencing is often important in the preprocessor. A model routine for rebuilding the base case helps manage sequencing. The first thing done when the user selects the menu option for rebuilding is a simple check of the data files to make sure that the country and region names are consistent and appropriate and that each region of the model has at last one country within it – in recent years, each region of the model has had exactly one country in it. The two primary files used are IFs.mdb and IFsHistSeries.mdb. The former file contains information on the regionalization, contains selected data such as population data by age cohort that are not easily organized by country and year alone, and serves as a storage location for the intermediate calculations of variables that are being processed in the preprocessor. The latter file contains the primary database of IFs and consists of tables by raw data series, all organized by country and year. A data dictionary file (DataDict.mdb) provides extended information, including sources, about each series in the raw data tables of IFsHistSeries.mdb.

A second preliminary step is the reading of population and GDP for each country/region. These two variables are the most fundamental for each country. The preprocessor requires them for weighting other variables when aggregating countries into regions. And the preprocessor often requires GDP per capita as an input to cross-sectional functions used for the filling of data holes. Thus every effort is made to assign values for these variables to each country in the basic data files and, for the very exceptional cases in which no values exist, the preprocessor assigned "1" to each (a GDP per capita of $1000) because it absolutely needs some values in order to proceed. It could be said about IFs that these two variables are the only completely essential numbers needed for the preprocessing and model calculations to proceed.

This second preliminary step of reading and processing population and GDP is actually done at the beginning of the population routine (which otherwise is documented in the next chapter). Specifically, population (POP) is read from the SeriesPopulationUNFilled table of the IFsHistSeries.mdb file and, via a call to ProcessEconomicWeights (in DataEcon) from DataPop, GDP is read at both market exchange rates (MER) and purchasing power parity (PPP). GDP at PPP is actually read in 1995 and 2000 dollars, so that both forecast series can be prepared. (See Section 5.1 on GDP at MER and GDP at PPP.)

The preprocessor then guides the process through the sequence of steps that are organized by substantive issue areas.

- It begins with population because that is the most self-contained area of the model, requiring only GDP from other areas. The preprocessor imposes total population data on the cohort-specific data by normalizing cohort numbers to the total. It reads values for a wide range of population-related variables: total fertility rate, life expectancy, HIV infection rate, literacy rate, etc. IFs uses cross-sectionally estimated relationships to fill holes in such data (generally with separate functions for the 1960 and 2000 data loads). Most often, functions driven by GDP per capita at PPP have had the highest correlations with existing data; the best functions have often been logarithmic, because the most rapid structural change occurs at lower levels of GDP per capita (Chenery 1979). The philosophy in demographics and in subsequent issue areas in the pre-processor is that values for all countries in IFs will come from data when they are available, but will be estimated when they are not.

- It proceeds with agriculture and energy, because the two partial equilibrium, physical submodels are closely linked to the general economic data, so that physical data for them must be carefully reconciled with the value data of the economic model. In agriculture, the pre-processor reads data on production and trade. It aggregates production of various crops into a single crop production variable used by the model. It similarly aggregates meat and fish production for the model. It computes apparent consumption. It reads data on variables such as the use of water and on the use of grain for livestock feed. It uses estimated functions or algorithms to fill holes and to check consistency (for instance, checking grain use against livestock herd and grazing land data).

- In energy, the pre-processor reads and converts energy production and consumption to common units (billion barrels of oil equivalent). It checks production and reserve/resource data against each other and adjusts reserves and resources when they are inconsistent. Null/missing production values are often overridden with a very small non-zero value so that a “seed” exists in a production category for the subsequent dynamics of the model (a technique used by the Interfutures model of the OECD). World energy exports and imports are summed; world trade is set at the average of the two and country-specific levels are normalized to that average.

- Flow then proceeds to the economic issue area. The outputs from processing of agricultural and energy data become inputs to the economic stage of pre-processing. The economic processing begins by reading GDP at both exchange rates and purchasing power and saving the ratio of the two for subsequent use in forecasting. The first real stages of economic data pre-processing center on trade. Total imports and exports for each country are read; world sums are computed and world trade is set at the average of imports and exports; country imports and exports are normalized to that global average. The physical units of agricultural and energy trade are read and converted to value terms. Data on materials, merchandise, service, and ICT trade are read. Merchandise trade is checked to assure that it exceeds food, energy, and materials trade, and manufactures trade is identified as the residual. All categories of trade are normalized. When this process is complete, the global trade system will be in balance. The use in IFs of pooled trade rather than bilateral trade makes this easier, but a similar process could be used for bilateral trade with Armington structures.

- The processes for filling the SAM with goods and services production and consumption, and with financial flows among agent-classes follow next. Stocks are computed as possible and are important to the long-term dynamics of the model itself.

- The sequencing of preprocessing thereafter is less important. The flow proceeds to the systems of data for education, domestic socio-political processes, international conflict, state failure, and values.

The remainder of this documentation will proceed in the same order as does the preprocessing.

2. Population

The first substantive issue area that the preprocessor addresses, largely because little is needed from other substantive areas, is population. As indicated in the previous chapter, the routine (called DataPop) begins by reading and storing population and GDP for use later in the routine and throughout the preprocessor.

The remainder of population data processing divides fairly cleanly into two steps: processing the population cohort data and processing assorted population variables.

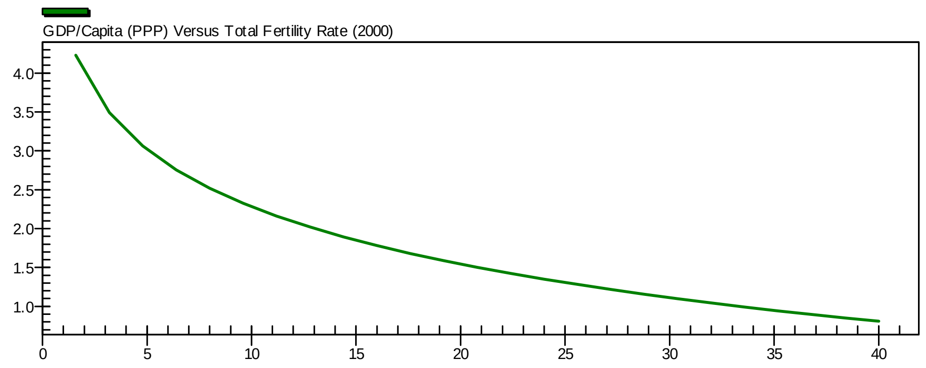
2.1 Cohort Data Processing

Cohort data are available in IFs.mdb for population by age and sex up to 100+. They are also available for fertility by age of the mother and for mortality, in the form of survival tables showing the number of an initial population that survive at each age, again through 100+. The raw data come from the United Nations Population Division and are revised on roughly 2-year cycles (such as the 2002 and 2004 Revisions). They are manually processed somewhat before being put into the IFs.mdb file. In particular, missing data are filled with cohort data from similar countries. For instance, data from China was used to create cohort structures for Taiwan.

The actual normalization of cohort population numbers to the total population of specific countries (or regional aggregations of countries) is done in the first year of the model itself. The value of this is that the user can change a population value and the cohort-structure will automatically adjust in the run of the model.

2.2 Aggregate Population Variables

The remainder of the routine reads and fills holes in a wide variety of demographic variables. For instance, total fertility rate (TFR) is read from SeriesTFR in IFsHistSeries. If the rebuilding process is creating a historic base case with initial year of 1960, the earliest year available is read (normally that will be 1960). If the rebuilding is of the 2000 base load, the value for 2000 is read if available, and the most recent or latest value is read if it is not. Cross-sectional functions based on GDP per capita (PPP) are used to fill holes when no value exists in the data set. For instance, the function below is used for the 2000 data load.



[Note DataPop is using 1995 – switch to 2000; many others need update; there is no temporal variation on function for labor participation rate]

Variables and the data tables from which they are read are:

Total fertility rate (SeriesTFR, GDP/Capita (PPP) Versus Total Fertility Rate (1995))
Infant Mortality (SeriesInfMort, GDP/Capita (PPP) Versus Infant Mortality (1995))
Life Expectancy (SeriesLifExpect, GDP/Capita (PPP) Versus Life Expectancy (2000))
Contraception Use (Series PopContrUseTotal%, GDP/Capita (PPP) Versus Contraception Use (1990))
Illiteracy/Literacy (Series Illiteracy %, GDP/Capita Versus Illiteracy (1995))
Calories consumed per capita (SeriesCalPCap, GDP/Capita (PPP) Versus Calorie Demand (1999))
Malnourished children (SeriesMalnChil%WB, Calories per Capita Maln Children % (1998-2000))
Malnourished Population percentage (SeriesMalnPop%, Calories per Capita Versus Maln Population % (1998-2000))
Population migration (SeriesPopMigration, algorithm)
Labor (SeriesLabor, GDP/Capita Versus Labor Participation Rate)
Female labor percentage (SeriesLaborFemale%, GDP/Capita (PPP) Versus Female Share of Labor Force (1995))
Crude birth rate (SeriesCBR. GDP/Capita (PPP) Versus Crude Birth Rate (1998))
Crude death rate (SeriesCDR, GDP/Capita (PPP) Versus Crude Death Rate (1998))
HIV infection rate (SeriesHIVRate,.01)
HIV infection peak year (SeriesHIVPeaks, 2100)
AIDS death rate (SeriesAIDSDths)
Urban population (SeriesPopulationUrban, GDP/Capita (PPP) Versus Urbanization Percentage (1998))

Literacy rather than illiteracy is used in the model so a simple conversion of 100-illiteracy is done.

A check is done to make sure that the percentage of those malnourished in the total population is not more than 70 percent of the number of children malnourished. This could happen is countries for which, for example, data on child nutrition were available but those for the entire population were estimated by function.

[the code form Mohammod on calories per capita and malnutrition for the MDGs needs to be documented]

The treatment of migration rates is more complicated than most of the series above. They are read from United Nations data not only for initial years but for the entire historic period, both to compute average patterns and to calculate an estimate of migrant population stocks from the integration of historic flows. When values for the forecast base load are read, historic inward flows to countries with GDP per capita less than $1000 are assumed to be extraordinary (such as flows into Eritrea associated with civil war) and are not used to initialize forecasting values.

When all country values for migration are read, they are balanced globally by summing across regions, using the average of inward and outward flows as a global total, and normalizing country values.

The growth rate of HIV infection is calculated by using data from 1999 and 2001 and computing a compound growth rate. Estimates of the peak year for HIV infection are taken from UNAIDS sources, as it the possible peak level in that year. The growth rates are checked against the peak year of infection rates and the peak level in that year. If the peak year has already passed and the growth rate of infection is not negative, it is changed accordingly. If the peak year is in the future, the rate of increase is checked against the estimates for the peak year and adjusted accordingly. The initial value of the AIDSdeath rate, needed in the model for 2000, is averaged from values in 1999 and 2001.

Urban population growth rates for runs beginning in 1960 are calculated from the 1960-1970 period. Those for 2000 are calculated from the 1990-2000 period.

After all of the cleaning and processing of demographic data, they are saved in a table of IFs.mdb called PopOutput. That table is used by the RebuildBase routine to create regional values (if countries are aggregated into regions) and to fill the base run file of the model.

3. Agriculture

The major tasks of the preprocessor in agriculture are to compute initial conditions for production, trade, and consumption in crops, meat, and fish and to compute initial conditions for the resource base for their production, especially land but also water and road infrastructure. In general, consumption is not available in databases and is computed as apparent consumption, that is as production plus imports minus exports. We follow that procedure in IFs.

A small number of parameters are read on a global basis (from IFs.mdb) before beginning the processing of country-specific initial values. These are ocean fish catch (OFSCTH), the annual slaughter rate (SLR) of animals from livestock herds (ideally these should be related to the character of the livestock herd by country with higher rates for poultry than for beef) and the amount of land withdrawn (LDWF) from agricultural use with the growth of population (this should ideally also be a country-specific specification).

3.1 Fish and Meat

The next steps involve the calculation of fish and meat production, trade, and consumption. IFs divides fish into aquaculture and marine catch categories. Fresh catch could also be conceptualized and forecast separately, but it is relatively small and is added to aquaculture.

Aquaculture (AQUACUL) is thus initialized as the sum of three variables, shown with associated data table and minimum default specification:

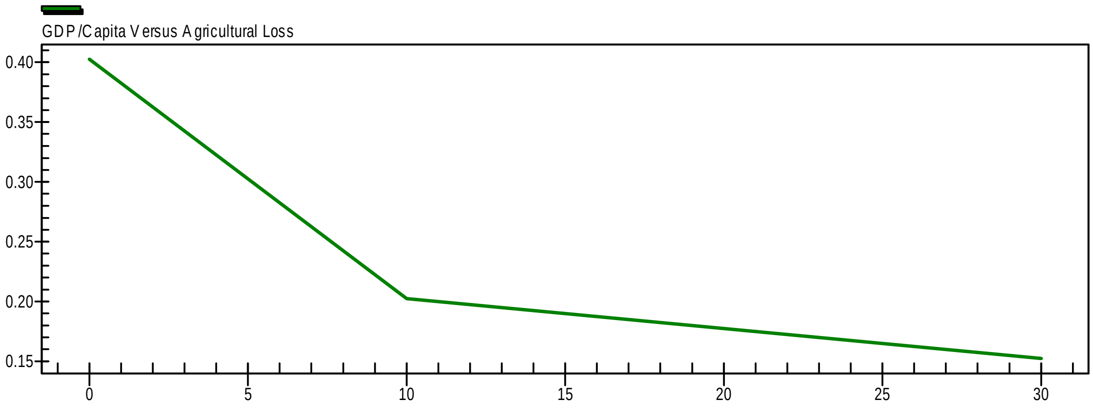
Marine aquaculture (SeriesAgFishAquaMarine, 0.001)
Inland aquaculture (SeriesAgFishAquaInland, 0.001)
Freshwater catch (SeriesAgFishFreshwaterCatch, 0.001)

Country-specific shares in marine fish catch are computed by reading the values in SeriesAgFishMarineCatch, specifying a minimum value of 0.0001 (this could be truly zero in land-bound countries) and normalizing the sum of country values to 1.0. In forecasting these shares of global catch can be applied to the global specification of total ocean fish catch (OFSTCH).

Marine catch (SeriesAgFishMarineCatch, 0.0001)

Total fish availability before trade (FISHAVAIL) is the sum of aquaculture and country share of ocean catch.

As the preprocessor moves next to meat and crop values, it begins by calculating production losses before consumption (for which there are very few data). Some food production will never make it to markets, but will be lost in the field or in distribution systems to pests, spoilage, etc. That loss (LOSS) is a function of GDP per capita in a table function (GDP/Capita Versus Agricultural Loss) that captures the tendency for the rate of loss to decrease with higher income levels. Loss rates for meat and fish are arbitrarily set at one-half those of crops. There is much arbitrariness in these specifications, but that is preferable to not treating loss at all, because reduction of rates of loss is an important driver of increases in calorie availability in dynamic forecasting.



Meat production is tackled next. Ideally apparent consumption should be production plus imports minus exports, but there is much missing data. In particular, we have no data on trade in fish.

[Note in preprocessor says there are now data on fish exports and imports – should use – consider putting notes like this into footnotes unless dealt with in first revision]

Therefore the preprocessor first calculates a default value for per capita consumption using the function GDP/Capita (PPP) Versus Meat Demand and multiplies it by population to give total country default consumption of meat and fish (PREDCONMEATFISH). Expected or predicted consumption of meat is that value minus fish availability. The assumption is made that at least 10% of the expected total consumption of meat and fish will consist of meat, so if fish availability exceeds 90% of predicted total consumption, the remaining fish is assumed to be exported. Similarly, if the fish availability meets less than 40% of total expected meat and fish production, the remainder is assumed to be imported. [This seems a rather high assumption of fish in the diet.]

Meat imports (SeriesAgMeatIm, 0.00005*GDP)
Meat exports (SeriesAgMeatEx, 0.00005*GDP)

Meat imports and exports are read from SeriesAgMeatIm and SeriesAgMeatEx respectively. If values are null, they are set at 0.0005 times GDP, approximately global average levels.

Meat production (SeriesAgProdMeat, algorithm)

The next step is reading of meat production where available from SeriesAgProdMeat. If values are null (or below 0.001), production is set at .001 and exports are set to production plus imports minus predicted meat consumption. Otherwise, we know meat production so can use it to recompute exports if those appear to be weakly specified in the data. Specifically, if meat exports are less than .0001 (likely indicating that zeros are specified in lieu of nulls), then they are recalculated as meat production plus imports minus predicted meat consumption.

Size of the livestock herd is calculated at meat production divided by the slaughter rate.

When the above processes are complete, there is an internally consistent data load for meat and fish production, trade, and consumption. It will still be necessary to normalize global exports and imports in meat and fish so that global sums are equal; the sums are computed at this point and the normalizations are done simultaneously with those for crops later in the preprocessor.

3.2 Crops

Turning attention to crops, total crop trade is summed from the trade of cereals, roots, pulses, and fruits/vegetables. In the case of production, data for fruits and vegetables are separated and production is the sum of five categories.

The preprocessor reads and cleans trade data first. It reads from the data tables indicated below and sets null values to roughly the global averages, computed as a portion of GDP. There are no data on trade in roots, and currently both imports and exports are set at 0.001.

Imports of cereals (SeriesAgCerealsIm, 0.005*GDP)
Imports of pulses (SeriesAgPulsesIm, 0.001*GDP)
Imports of fruits and vegetables (SeriesAgFruVegIm, 0.001*GDP)
Exports of cereals (SeriesAgCerealsEx, 0.005*GDP)
Exports of pulses (SeriesAgPulsesEx, 0.001*GDP)
Exports of fruits and vegetables (SeriesAgFruVegEx, 0.001*GDP)

The above exports and imports are summed globally in preparation for normalization to a single global sum for trade in each category.

Moving to production, the variables read and the values used to replace nulls are:

Production of cereals (SeriesAgProdCereals, 0.005*GDP)
Production of roots and tubers (SeriesAgProdRootsTub, 0.00025*GDP)
Production of pulses (SeriesAgProdPulses, 0.0001*GDP)
Production of vegetables, melons (SeriesAgProdVegMel, 0.0005*GDP)
Production of fruits except melons (SeriesFruitsExclMelons, 0.00002*GDP)

Crop production is summed from the above five categories.

The rate of growth in crop production (AGPGR) is also needed later in the agricultural preprocessor to compute growth in yield. That is obtained from looking at the period after 1980. Data for different crop types and for different countries varies in availability. So the computation is fairly complicated in terms of specifying end years, but in all cases the intention is to obtain an annual compound growth rate from as many years post 1980 as possible. For those countries where data is simply not available, a function is used to estimate a value – based on empirical analysis, the function assumes that growth in the least developed countries is about 2.7% per year and that it drops to about 1.5% per year for the most developed.

At this point trade in meat, fish, and crops are normalized so that global exports and imports will balance.

When the above block is complete, the preprocessor has production and trade data for initialization, but no check has been undertaken to assure that apparent consumption is reasonable or that the cereal needs for industrial and livestock feed purposes are consistent.

Feed demand can rely on data:

Feed demand from grain (SeriesAgGrainLiv%GrainCon, GDP/Capita (PPP)
Versus Grain Feed as % of Total Grain Use (1995))

Consistency checking on apparent consumption in cereals is done in several steps. First, an apparent consumption figure is computed (AppConCereals), applying the loss factor to production: production minus loss plus imports minus exports. Loss-adjusted production plus imports minus exports obviously must exceed zero.

The percentage value of feed demand for grain is applied to apparent grain consumption to calculate feed demand.

Industrial demand for food (IndDem) generally is computed as an arbitrary 10% of apparent production of crops.

At this point food demand is available as a residual, but it can be highly variable by country (most counties range between 0.275 and 0.5 metric tons/capita). Although data for agricultural production, imports and exports, and even feed demand are not bad, the values for losses of food produced and industrial demand are fairly arbitrary. They are therefore adjusted as necessary to bring food demand into, or at least near to that normal range.

Apparent food demand per capita (AppFDemPC) is calculated as production after losses plus imports and minus exports, industrial demand and feed demand, all divided by population. If it is less than 0.2 metric tons, crop losses are first reduced to as little as 10 percent. Apparent food demand per capita is recomputed and if it is still less than 0.2 metric tons, industrial use of food is reduced to as little as 1 percent of the initial calculation. Apparent food demand per capita is again recomputed. If the number is less than 0.05 metric tons, an error flag is given to the user.

When apparent food demand per capita exceeds 0.6 metric tons, the process is reversed. Losses are increased to as high as 50 percent. If apparent demand is still too high, industrial demand is increased to as much as 3 times the initial value (that is, up to 30% of apparent production of crops).

3.3 Production Resources

Processing then moves to land use. The following series are read:

Land area total (SeriesLandTotal, sum of agricultural land types)
Crop land (SeriesLandCrop, based on AGPTot)
Grazing land (SeriesLandGrazing, based on total land area)
Forest area (SeriesLandForest, based on total land area)
Other land (SeriesLandOther, based on total land area)
Build area (SeriesLandUrban&Built, GDP/Capita (PPP) Versus Land/Capita
Built Up (1992-93))

Null values for crop land, or values less than .01 are filled with .5*AGPTot, meaning that 2 tons/hectare is assumed average yield for countries for which there is no data, most if which will be less developed. The value for crop land is increased if the apparent yield would be more than 10 tons/hectare. These changes obviously privilege production data over land data, which is assumed to be of lower quality.

When grazing land data are missing, the value is set to a global average of 22% of land area. Similarly when forest area or other use is missing, they are set to 30% or 25% of total land area, respectively. Urban/built area is read of filled via function on a per capita basis (GDP/Capita (PPP) Versus Land/Capita built Up (1992-93, Linear) or GDP/Capita (PPP) Versus Land/Capita Built Up (1992-93, Logarithmic)), so total urban area requires multiplication by population and division by 1000 to convert to million hectares. Because data on other land area actually include urban and built land, the other category is reduced by the urban/built value. [no reduction is made if urban exceeds other; should be substantial reduction leaving residual]

In order to initialize growth in agricultural production in early years, the model requires values for both the growth in yield and the growth in land devoted to crops. If data exist for land under crops in 1992 and 2001, the initial target growth rate in crop land (TGRLD) is computed as the compound annual growth rate over that period. If data do not exist, a maximum likely rate of growth in crop land (MaxGrow) is computed as 1.5 percent minus a factor related to the extent by which existing crop land exceeds 20 percent of total land. That is, the assumption is that potential for growth drops as the crop land portion increases. An actual estimate for growth in crop land is calculated with the knowledge that the recent historic rate of growth in the poorest countries has been about 0.9 percent and that in the richest countries has been about -0.2 percent. So for the countries without land data the estimate is based on a linear function of GDP per capita between these two values, bounded by the maximum number indicated above. The final value is bound between -0.3 percent and 1 percent.

When the agricultural production data were read, a value for growth was calculated or estimated for missing data (AGPGR). The difference between that value and the growth in land is the basis for the initial target growth in yield (TGRYL). The final value is bound between 0.7 percent and 3.5 percent.

One additional initial condition is set, that for the share of agricultural investment going to land (IALK), the residual going to improvements in capital that affect yield. It is set at 0.25 or 25 percent. Perhaps this should vary with the balance between TGRLD and TGRYL, but land requires various kinds of investment even when the amount devoted to crops is stable or decreasing.

Because at this stage of the preprocessor land data have been processed, it is possible to make a refinement in the amount of grain going to livestock. Given the size of the livestock herd (LiveHerd), it is possible to compute the likely amount of total feed (from land and supplementary) needed for them, based on a cross-sectionally estimated function of feed demand per unit (GDP/Capita (PPP) Versus Feed Requirements). It can be assumed that the grazing land can support at the very minimum 0.01 ton of feed equivalent per hectare. If this minimal level of productivity, multiplied by the size of the herd and subtracted from total feed needs, is less than the feed demand estimated earlier from a function, that feed demand is reduced. This is another of example of attempting to privilege data relative to functions.

The final two sets of calculations in the preprocessor for agricultural data establish values for two infrastructure variables, road networks and water resources. Data come from series:

Road network (SeriesRoadsTotalNetwork, “GDP/Capita (PPP) Versus Road Network/Land Area (1999))
Renewable water (SeriesWaterAnRenResources, 0.1)
Water withdrawals (SeriesWaterAnWithdrawals, based on water use)

Road density is computed by reading the total size of the network and dividing it by total land area. In the case of nulls, an estimated function directly calculates a value for density from GDP per capita.

Water use per capita is computed by reading total water use and dividing it by population. In the case of nulls, an estimated function directly calculates a value for it from the size of agricultural production per capita. [it looks like the function input should be AGPTot/pop but it now is AGPTot]

Total water resources are read when possible, but in the case of null values it is assumed that they are ten times the total water use (per capita use times population).

The penultimate step in the agricultural data preprocessor is the writing of all values to the AgriOutput table of the IFs.mdb file.

In the process of rebuilding the base, a final subroutine aggregates country-specific data into regions and prepares the final data load for the model. Currently, the numbers of countries and regions are identical so that no aggregation is actually needed.

4. Energy

The major tasks of the preprocessor in energy are: to compute production in multiple energy categories (oil, natural gas, coal, hydroelectric, nuclear, and other renewable) as well as to compute known reserves and ultimately recoverable resources in the same categories; to compute trade in energy as a single category; to compute apparent production of energy and to assure that it is reasonable relative to the size the economy. In general, consumption is not available in databases and is computed as apparent consumption, that is production plus imports minus exports. We follow that procedure in IFs. In the energy model, billions of barrels of oil equivalent are used as a standard common unit of measurement for volumes.

A single parameter is read on a global basis (from IFs.mdb) before beginning the processing of country-specific initial values. It is the price of oil per barrel in the initial year of data (WEPBYYEAR), needed to compare/reconcile physical and monetary variables.

In addition, parametric values for possible maximum production by country of oil, gas and coal (ENPOILMAX, ENPGASMAX, and ENPCOALMAX) are read with null values being set to 0 or no limit.

4.1 Production, Consumption, and Trade

Following these readings of parameters, a long series of data series around production, consumption and trade are read:

Production of oil (SeriesEnProdOil, none)
Production of oil, BP data (SeriesEnProdOilBP, algorithm)
Production of gas (SeriesEnProdGas, none)
Production of gas, BP data (SeriesEnProdGasBP, algorithm)
Production of coal (SeriesEnProdCoal, none)
Production of coal, BP data (SeriesEnProdCoalBP, algorithm)
Consumption of electricity (SeriesEnConElec, algorithm)
Thermal share of electricity production (SeriesEnThermalElec, algorithm)
Consumption of hydro (SeriesEnConHydro, none)
Consumption of hydro, BP data (SeriesEnConHydroBP, algorithm)
Production of nuclear (SeriesEnProdNuclear, none)
Consumption of nuclear, BP data (SeriesEnConNucBP, .00001)
Production of geothermal (SeriesEnProdGeoTherm, algorithm)
Production of solar (SeriesEnProdSolar, .000001)
Production of tides/wave energy (SeriesEnProdTideWave, .000001)
Consumption of photovoltaic (SeriesEnConPhoto, .00001)
Consumption of wind (SeriesEnConWind, .000001)
Energy exports (SeriesEnExports,algorithm)
Energy imports (SeriesEnImports, algorithm or "GDP/Capita (PPP) Versus
Energy Demand per GDP Dolar (1995))

Oil exports (SeriesEnExportsOil, none)
Oil imports (SeriesEnImportsOil, none)
Electricity consumption per capita (SeriesEnElecConsPerCap, GDP/Capita (PPP 2000) Versus Kilwatt-Hours per Capita (2000) Linear)

Electricity consumption per capita is read and nulls are filled with the function indicated. These values are not used elsewhere in the energy preprocessor and are computed for the purposes of filling data about electrical infrastructure development for use in the economic model.

Because there is limited data on total energy consumption, the model relies heavily on apparent consumption, computed as production plus imports minus exports. Thus the first major task of the preprocessor is the reading and cleaning of production and trade data.

There are two major sources of energy production data. The first is British Petroleum, which has the most up-to-date data, but also tends to restrict coverage to the largest countries. The second is the International Energy Agency/United Nations, with data that tend to be slow in coming but have extensive country coverage. The philosophy in the preprocessor is to look to the British Petroleum data first and, when nulls are found, to look next at the data from international organizations. There are also occasions when zero values are found in some of the data; these are assumed to be nulls rather than true zeros.

When only nulls (or zeros) are found in both data sources, a low value is put into the production variable. On a global average oil production in billion barrels is about 0.008 times GDP in billion dollars and natural gas production is about 0.005 times GDP. Values of 0.0002 times GDP are used to replace zeros or nulls. These are, in a real sense, “starter” values for production in these cases – it allows the model to ramp up production if data show reserve availability and simply to keep it at a minimal level is there are no reserves to support production. Similarly, nulls or zeros in coal production are replaced by 0.0005 times GDP, and electricity production is set at 0.0002 times GDP, as is hydroelectric production. In the case of the small energy components such as nuclear and geothermal production, nulls are replaced with the absolutely small values of 0.0001 billion barrels of oil equivalent. In the case of even smaller energy components such as photovoltaic, wind, solar (such as solar towers), and tide/wave production, nulls and zeros are replaced with absolute values of 0.000001 billion barrels of oil equivalent. A temporary sum (ENPTempTotal) of all production is computed.

Energy imports (ENM) and exports (ENX) are computed next. When values are found to be null, data tables for imports and exports oil alone are scanned and used, on the assumption that oil accounts for most energy trade. If no values are found in either source, energy exports are set at 0.0002 times GDP. In the case of null imports, a very conservative total energy demand is calculated at 0.0002 times GDP (the world average is about 10 times that at 0.0024 times GDP), and a very conservative level for imports is set at the conservative energy demand value minus the temporary calculation of production

(ENPTempTotal) and plus exports.   As one last constraint on imports, the value of them (physical imports times world energy price) is constrained to be no more than 10 percent of GDP.

The above description of the inputting of energy production and trade has been based on the procedure for the preprocessor's handling of data set up for normal forecasting. When forecasting is being undertaken over a historical period (starting in 1960) there are additional complications introduced by the shortage of energy data at the beginning of that period, particularly those for energy production. Thus more weight is placed on energy trade and on reasonable estimates of energy demand. Specifically, energy imports (ENM) are read first, filling null values with a cross-sectionally estimated function of GDP per capita. Then a reasonable approximation of energy demand (ApproxEnDem) is calculated at 0.002 times GDP. An estimation of a minimum for local production (LocalProdMin) is set at 0.15 times the approximation of energy demand and an estimate for local production (LocalProdEst) is set at the maximum of that minimum or approximate demand minus imports (which is more likely the maximum). In the replacement of null production values for the historic data load, this estimate of local production is used as a base. Null oil and gas values are replaced by 0.1 times the local production estimate and null coal values by 0.2 times the estimate. Null values for hydroelectric production are set at 0.01 times the estimate. Null geothermal and nuclear values are put absolutely at 0.000001 billion barrels of oil equivalent and null values for photovoltaic, wind, solar, and tide/wave energy, as technologies that did not exist in 1960 and were insignificant through 2000, are set at zero.

After basic production and trade data are read for both historic and future data loads, the "new" renewable energy forms of geothermal, photovoltaic, solar, wind, and tide/wave are summed into an other renewable energy category for the initial data load.

In addition, the preprocessor sums global energy imports and exports and normalizes country values to the average of global sums of the two.

At this point the flow of the preprocessor moves to the consumption side and computes apparent consumption (AppCon) as the sum of production plus imports and minus exports.   For the purpose of reference, a global ratio of apparent consumption to global GDP is computed (AppConF).

Because of the poor quality of data in many countries, a number of checks are then undertaken on apparent consumption. A cross-sectionally estimated function, "GDP/Capita (PPP) Versus Energy Demand per GDP Dollar (1995)" provides a reference value of energy consumption per unit of GDP (CompEnDemPerUnit) that is multiplied by GDP to obtain a computed value of energy demand (ComputedEnergy).

If the apparent energy consumption is negative or if it is less than 40 percent of the computed value, it is assumed that the data on production and/or imports of energy were too small. Apparent consumption of energy is set at 40 percent of the computed value (based on the cross-sectionally estimated function) and all energy production values and adjusted upward proportionately. The apparent energy consumption value is recomputed.

[This next block appears unnecessary, perhaps a leftover; put a breakpoint inside the block to see if ever accessed.] It is possible that even with the upward adjustments of production, apparent energy consumption is still too low. If the computed energy from the cross-sectional function is more than four times apparent energy consumption, then energy production is again increased with an adjustment factor.

It is also possible that production or import values from data can be too high, so that apparent energy consumption relative to the computed value based on the function is unreasonably large. A check is made to see if apparent values are more than three times computed ones, or more than six times in the historic load (because communist countries like China were incredibly energy inefficient in the 1960s). If apparent values are that much too high, the assumption is that exports are probably too low in the data load and the preprocessor adjusts them upward accordingly, but does not allow exports to be more than 95 percent of the sum of oil and gas production. A further check is undertaken inside this adjustment process. If exports minus imports in value terms are more than 70 percent of GDP, then the surplus above that 70 percent is used to reduce production of oil and gas accordingly and thereby to bring exports back down. This adjustment was introduced in part because of the data for Azerbaijan. Oil exports have growth so rapidly that GDP data have probably not keep up with them (or perhaps the exports have been shipped at less than global oil prices). It would have been possible and perhaps even preferable to increase the GDP data, but those have been privileged in the data set because they tend to be better than energy data. Thus production and exports of oil for that country are adjusted by the preprocessor.

Because of possible adjustments to trade in the process of comparing apparent production with reasonable ranges of it, global imports and exports are again summed and normalized.

4.2 Reserves (with Production Constraints) and Resources

The preprocessor moves next to data for known energy reserves (as opposed to ultimately recoverable resources). It reads series for the major fossil fuels and one that estimates the upper limit for hydroelectric capacity:

Oil reserves (SeriesEnReserOil, .001)
Gasreserves (SeriesEnReserGas, .001)
Coal reserves (SeriesEnReserCoal, .01)
Hydroelectric reserves (SeriesEnReserHydro, .01)

In the process of reading the data, nulls are filled with absolute values as indicated above. Reserves of coal and hydroelectric capacity are converted to billion barrels of oil equivalent. Reserves of oil and gas are set at least to 10 times current production levels, reserves of coal are set to at least 30 times production, and reserves of hydro are set to at least 1.5 times production.

The next step is to compute production growth rates for the initialization of such growth in the model. For the values in the forecast runs, oil production data from British Petroleum in 1998 and 2001 are used to compute an annualized growth rate. If either of those values are null or zero, the growth rate is set at 5 percent, but is reduced (to as low as 1 percent) as the reserve to production ratio declines. Gas, coal, and hydroelectric production growth rates are based on values in 1995 and 2001, similarly dealing with missing data. The reason for looking at a shorter period for oil is that oil production has been topping in more countries. Nuclear production growth rates look only at 1999 and 2001 because policies on nuclear energy are subject to more recent change. Other renewable components are examined across the most five year period for which data are typically available. If there are missing data in any of the components, the overall growth rate for other renewable energy is arbitrarily set at 4.5 percent.

The process for computation of production growth rates in the 1960-based historic run uses data over longer periods when they are available. Oil data from British Petroleum are examined over the 1965-70 period that initiates that database if they are available; otherwise the preprocessor looks to data from the UN over the 1960-65 period. Similarly, natural gas data come from British Petroleum for 1970-80 when available, otherwise from the UN for 1960-70. Coal data come from British Petroleum for 1981-1990 or from the UN for 1960-70. Nuclear data come from British Petroleum for 1970-2000 or from the UN for 1960-65. Hydroelectric data come from British Petroleum for 1965-2000 or from the UN for 1971-2000. British Petroleum data are privileged for their reliability, even when the period they cover does not go back to the 1960 starting data of the historic run.

Having completed the processing of data on known reserves, the preprocessor turns to ultimately recoverable resources. The first step is reading several data series:

Oil resources from the USGS (SeriesEnResorOilUSGS, algorithm)
Natural gas liquid resources from USGS (SeriesEnResorNGLlUSGS, algorithm)
Gas resources from the USGS (SeriesEnResorGasUSGS, algorithm)
Oil resources from the WEC (SeriesEnResorOil, not used)
Gas resources from the WEC (SeriesEnResorGas, not used)
Coal resources from the WEC (SeriesEnResorCoal, algorithm)
Synthetic energy resources from the WEC (SeriesEnResorSynthetic,)

At this point the preprocessor has been switched entirely from WEC values for oil and gas to USGS values. Values for oil and natural gas liquids are read, converted from million barrels of oil equivalent to billion barrels, and summed because the model, as do most others, treats them as a single category. If data for oil are missing, the resource value is set at 1.6 times the reserves, and missing data on natural gas liquids are set at zero. In addition to these estimates of undiscovered resources, the USGS and other sources presume that currently discovered reserves can be extended with new technology. The amount of that potential extension is assumed to be 70 percent of existing reserves and that amount is added to the total for ultimately recoverable but not currently identified resources. That value is presumed to be at least twice known reserves and adjusted upward accordingly if not.

With respect to natural gas, data are again taken from the USGS source. Values are converted to billion barrels of oil equivalent and missing values are set at 1.6 times reserves. Reserve extensions are assumed to be 40 percent of exiting reserves and ultimately recovered resources are the sum of the two values but not less that 1.6 times known reserves.

Coal resources are read from WEC data and converted to billion barrels of oil equivalent. Missing or exceptionally low values are set at twice known reserves and values read are adjusted upward to at least 30% above known reserves if they are lower than that in the data file.

Data for synthetic fuels (basically tar sands, oil shale, and very heavy oil) are read from tables based on WEC data. Missing values are set to only 0.001 billion barrels because only a limited number of countries have such resources.

There also needs to be some upper limit on renewable energy resources and data on such limits (that is, estimates of them) is extremely scarce. The preprocessor looks to total land area for the foundation of an estimate. It reads that area (LandArea) from the IFs.mdb table to which the agricultural preprocessor routine wrote it. The ultimately recoverable (maximum annual production value) for renewable energy is assumed to be a function of the average of the population and the land area.   The logic behind this crude formulation is that it should go up with land area for some countries (Saudi Arabia but not Greenland) and that the population helps indicate better climate for harvesting renewable resources. This average value is divided by eight as a scaling factor to bring it in line with subjective reading of literature in the area. The entire procedure is, of course, highly subjective and arbitrary. At this point the model pays very limited attention to such constraints in its forecasting.

For the historic (1960) load only, the values for oil, gas, and coal resources are increased by the sum of production between 1960 and 1995. The logic of this is straight-forward: recent estimates of ultimately remaining resources exclude the production of those years and that increment was, in fact, available in 1960.

4.3 Parameters and Wrap-Up

The final calculations of the preprocessor set some parameters, as opposed to initial conditions. Values are specified for the capital-to-output ratios (QE) of all energy forms. For coal, hydroelectric, nuclear, other renewables, and synthetics, these values are set at 90, 180, 160, 150, and 170, respectively. These forms are assumed to have relatively consistent costs across countries and regions. In contrast, the capital costs of oil and gas production vary dramatically across countries, with much lower extraction costs in Saudi Arabia, for example, than in the United States. The base values for countries with very high reserves are set at 60 for both oil and gas. Those values are reduced for countries with reserve to production ratios that exceed 15 years.

Another parameter set is the ratio of reserves to production below which production is assumed to begin to face physical constraints that reduce it (PRODF). For oil, gas, and coal that parameter is set to 15. For countries with ratios of coal reserves to production that exceed 200, however, it is assumed that some kind of restraint on production already exists, perhaps in the safety or environmental issues surrounding coal extraction and/or use. The ratio for the parameter that determines production constraints is set at 100. The ratio for hydroelectric production is set at 2.0, because the “reserve” constraint here is one of available sites, not physical resources.

[there is also code for the setting of rdm for energy types; it apparently is irrelevant because the value of that multiplier is set at 1.0 for all countries and types; it should be commented out or removed.]

The remaining code of the preprocessor simply processes the values set in that described above, writing values to appropriate tables and files.

5. Economics

Preprocessing related to economics is the most extensive and complicated of all the preprocessing of data for initialization of IFs. It involves several important elements in roughly the following order:

1. Reading and processing of GDP at both market exchange rates (MER) and purchasing power parity (PPP). As noted in the discussion of population, there are only two numbers that are absolutely essential for the physical functioning of forecasting for any geographic entity in IFs: total population and total GDP at MER. Obviously, other numbers are required for meaningful forecasting, but the preprocessor will help the user identify those with initial "guesses" based primarily on cross-sectionally estimated functions. GDP at MER and PPP must themselves, however, be prepared for forecasting.
2. Reading and processing total country trade and initial aggregate current account information. Global trade and current account balances must be maintained so adjustments are made as necessary at this point.
3. Reading and processing the agricultural and energy trade that the preprocessor routines for agriculture and trade handled and balanced at the physical level. Converting the physical values to monetary ones and making sure that total trade numbers are consistent with them (that is, that total trade is greater than the total of them and that there is some "headroom" for other trade).
4. Reading monetary values for merchandise trade and service trade and reconciling them with total country trade; also reading and checking arms trade, manufactures trade, and ICT trade to assure that they fit within merchandise and services trade and they, too, are complete (filled by functions as necessary) and globally balanced. Trade in all sectors of the model must be both reconciled with country totals and balanced at the global level.
5. Processing other current account expenditures (in aggregate) and reconciling with GDP. Reading value added for the sectors of the economy where data are available, and estimating values for sectors such as energy and materials where they are not.
6. Reading IO data from GTAP, collapsing matrices to the sectors of IFs and creating a set of generic matrices related to development level for the process of filling country holes and preparing for structural change in forecasting.
7. Determining sectors of origin for household and government consumption and more generally fleshing out the final demand side of the economic system.
8. Moving to the broader social accounting matrix, beginning with government consumption, transfers, and revenues and proceeding to corporate and household flows.
9. Representing government and country-total liabilities and assets (stocks as opposed to the earlier flows).
10. Reading FDI and portfolio flow data and integrating over time to create stock values.
11. Reading and globalizing worker remittance data, which is available only for recipients.

12. Initializing World Bank and IMF flows and stocks.
13. Investigating major discrepancies in current and capital account information.
14. Initializing GDP Growth and Income Distribution
15. Adjusting Input-Output, Final Demand, and Value Added Data for greater consistency internally and with physical values of energy production and apparent consumption. An iterative process is used for these adjustments, generally privileging the IO data (except relative to the physical energy data), and most often privileging the value added data on manufactures and services relative to the final demand allocations by sector.
16. Computing regional values when the model uses multi-country regions (which it does not currently do).

There are a number of initial conditions specifications that are not geographically specific but are necessary for the functioning of the rest of the economic pre-processor. They are read early in the process from the IFs.mdb file, Global Parameters table:

The basic ICT share of the economy (ICTSHR)
World energy price, base year (WEPBYEAR)
The World Bank loan interest rate (XWBLINTR)
The base rate of repayment on World Bank loans (XWBLNREPR)
Annual growth in the World Bank's loan portfolio (XWBLOANR)
World Bank's loan to equity ratio (XWBLNEQR)
IMF credit interest rate (XIMFCRINTR)
The base rate of repayment on IMF credits (XIMFCRREPR)
Annual growth in IMF credits (XIMFCREDITR)
IMF's credit to equity ratio (XIMFCREQR)

From the multiple parameters file of IFs.mdb two additional parameters are read that are again global, but have dimensionality:

Crop prices, per ton (WORLDAP1)
Meat prices, per ton (WORLDAP2)

Finally, there are small number of variables needed in economics preprocessing that are read from the PopOutput table of IFs.mdb (they were read and cleaned in the population pre-processor):

Foreign-born population (Migrant stocks)
Crude birth rate (CBR)
Crude death rate (CDR)

5.1 GDP at Market Exchange Rates and Purchasing Power Parity

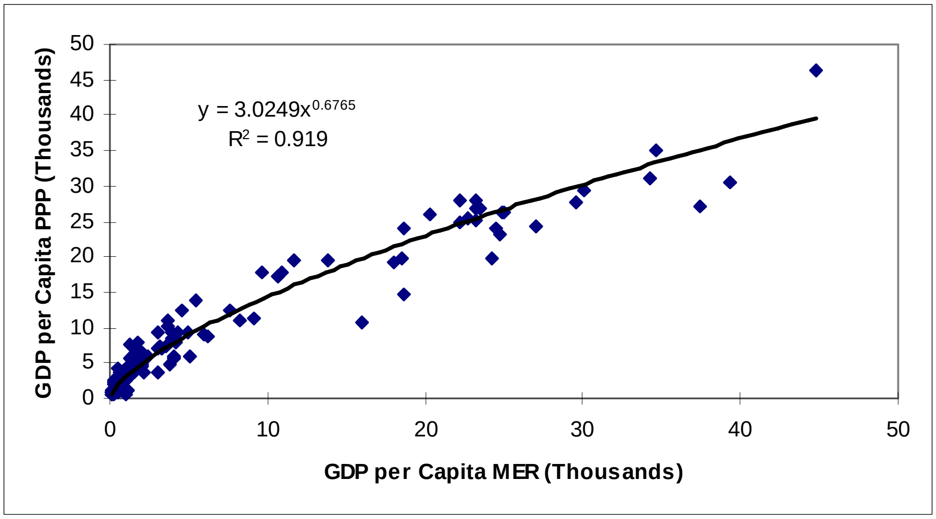
One issue that it was necessary to address on the front end of work in the economic model was whether to develop it at market exchange rates (MER) or at purchasing power parity (PPP). The model necessarily will be initialized and run in one or the other, with the secondary representation of economic variables being calculated from the dominant one. In the case of IFs, the decision was made to base the model at MER. MER data are the most extensive and are updated on a most timely basis. MER is also important for the representation of movements in exchange rates for equilibration of trade and current accounts.

At the same time, GDP at PPP quite consistently proves the better driver of many other variables within IFs, reinforcing the importance of the model's carrying information on both. For example, estimations based on historic data for social variables such as fertility rates, life expectancy, or democratization are stronger when based on PPP. Some variables, like energy demand, are less clearly related to one or the other.

The first step in the pre-processor is the reading of GDP at MER in year-2000 based dollars and the saving of it for future use. GDP per capita is computed from it (GDP2000PC MER). This is considered a required series for all countries, but if it is missing, values of $1000 per capita (and at least $1 billion for the total GDP) are entered). In addition values for GDP at MER in year-1995 based dollars are read and saved, also with conversion to GDP per capita. The purpose in the model of carrying these values based on $1995 is to facilitate back-conversion of PPP-based values such as GDP from 2000$ to 1995$ for provision of information to other modelling projects that use 1995$ for drivers.

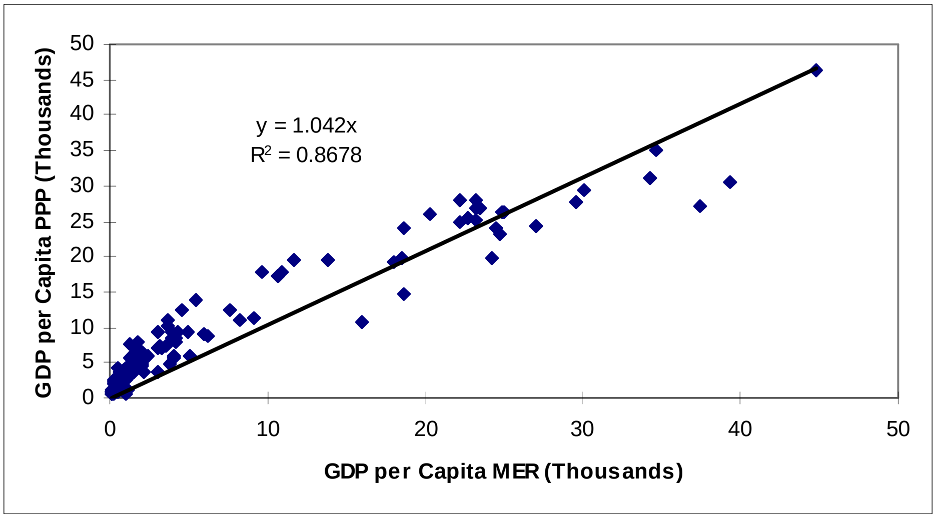
GDP at MER 2000$ (SeriesGDP2000, required)
GDP at MER 1995$ (SeriesGDP1995, required)
GDP per capita at PPP 2000$ (Series2000PCPPP)
GDP at PPP 1995$ (SeriesGDP1995PPPWDIFilled)

Next, values for GDP per capita at PPP for year 2000 (in current dollars so that 2000 values are in 2000$) are read. Although the data set is quite complete, the cross-sectionally estimated function below is used to fill any country holes (GDP/Capita Versus GDP/Capita at Purchasing Power Parity (2000) Power Function). In addition the values for GDP per capita at PPP for year 2000 are read and holes are filled with the cross-sectionally estimated function (GDP/Capita Versus GDP/Capita at Purchasing Power Parity (1995)). When these reads and the hole-filling are complete, there will be four complete sets of GDP per capita values for the 2000 base year: 2000$ and 1995$ at MER; 2000$ and 1995$ at PPP.



The United States is the numeraire country for PPP. That is, its GDP should be the same at MER and PPP. As a check on the above function, that is very nearly true (34.6 at MER and 35 at PPP – the actual cross-over point where MER=PPP is at 30.62). At the top of the graph the point is Luxembourg at 44.8 and 46.7, respectively.

The function above has a very high R-squared. The function below, using a linear formulation has a high one, but is quite clearly not as strong (both functions were forced through the origin).



Although it is not handled in the pre-processor, and information about it is not essential to understanding the functioning of the pre-processor, the relationship between GDP at MER and GDP at PPP during he forecasting process is so important to the model that it will be described here. The first (power) function above suggests that in forecasting there are two kinds of convergence to consider in longer-term forecasting:

1. The function indicates that, in general, the ratio between GDP per capita at MER and GDP per capita at PPP is closer to 1 in more developed countries than it is in less developed countries. That is, moving up the development curve of the function should gradually bring GDP per capita at PPP closer to values at MER, along the path indicated by the function (GDP/Capita Versus GDP/Capita at Purchasing Power Parity (2000) Power Function).
2. Although some countries are quite a distance from the function, there is substantial reason to believe that such discrepancies indicate errors in data or temporary variations from the function related to currency valuations or other forces. That is, most countries should, over time, gradually move the relationship between GDP per capita at MER and PPP towards that of the line.

First year. In the first year of the actual execution of the model, GDP at MER (2000$) is built up from the initial conditions of the expenditure components (as will be discussed later in this chapter) and GDP per capita is calculated through division by population.

GDP per capita at PPP in both 2000$ and 1995$ are read from files prepared in the pre-processor. This allows the computation of conversion rates from MER to PPP for both 2000$ and 1999$, respectively: PPPCONV=GDPPCP/GDPPC and PPPCONV95=GDPPCP95/GDPPCP.

The next step is to lay a foundation for the convergence over time of country-specific values to the functions that relate GDP/capita at MER to GDP/capita at PPP. One basic function is used to compute anticipated GDP/capita at PPP (2000$) from GDP/capita at MER (2000$): GDP/Capita Versus GDP/Capita at Purchasing Power Parity (2000) Power Function. It produces a computed value of GDP per capita at PPP ($2000) called GDPPCPComp. The ratio of the actual value to the computed value is saved as the FunctionRatioPPP = GDPPCP/GDPPCPComp.

Similarly, a function (GDP/Capita Versus GDP/Capita at Purchasing Power Parity (1995)) is used to compute the expected value of GDP/Capita at PPP (1995$), saving it in GDPPCP95Comp. Again a ratio of the actual value of the computed value is saved: FunctionRatioPPP95 = GDPPCP95/GDPPCP95Comp. In this case also the GDP per capita at MER in 2000$ (GDPPC) is used as input to the function.

In future years the two ratios of actual values to those on the functions will be allowed to converge over time to 1, gradually bringing country values at PPP onto the function linking MER and PPP.

Subsequent Years. In years beyond the first, GDP and GDP per capita are computed at MER in 2000$ in processes that are documented elsewhere. The variable is carried as GDPPotPC. Its values are again used for each country to drive the two functions that produce computations of GDP per capita PPP in 2000$ (GDPPCPComp) and 1995$ (GDPPCP95Comp), respectively:

GDP/Capita Versus GDP/Capita at Purchasing Power Parity (2000) Power Function
GDP/Capita Versus GDP/Capita at Purchasing Power Parity (1995)

Basic conversion ratios from GDP per capita at MER to GDP per capita at PPP are computed from them:

ConversionRatio = GDPPCPComp/GDPPotPC
ConversionRatio95 = GDPPCPComp95/GDPPotPC

For 2000$ the ratio of initial country values to the function (FunctionRatioPPP) is allowed to converge towards 1 over 100 years in the computation of a BaseRatio:

BaseRatio = ConvergenceOverTime(FunctionRatioPPP, 1, 100)

The actual conversion factor for MER (2000$) to PPP (2000$) is the product of the BaseRatio and ConversionRatio:

PPPCONV = BaseRatio*ConversionRatio

The total convergence of PPPCONV to 1 is limited to 5% per year. PPPCONV is used compute the GDP at PPP and GDP per capita at PPP (2000$):

GDPP = GDP * PPPCONV
GDPPCP = GDPPC * PPPCONV

For 1995$ the process is exactly analogous. Again, a BaseRatio is computed from initial conditions:

BaseRatio95 = ConvergenceOverTime(FunctionRatioPPP95, 1, 100)

The conversion factor for 1995$ is a product and change is again limited to 5% per year:

PPPCONV95=BaseRatio95*ConversionRatio95

The product is used to compute GDP at PPP and GDP per capita at PPP (1995$):

GDPP95 = GDP * PPPCONV95
GDPPCP95=GDPPC * PPPCONV95

The Use of GDP/GDPP as Drivers

Throughout the rest of the model, most often GDP at PPP is used as a driver of variables that depend on GDP, especially socio-political variables. In addition food demand uses GDP per capita at PPP and energy demand uses GDP at PPP.

5.2 Trade and Current Account, Other Expenditures, and Value Added

5.2.1 Basic Checks on Total Trade and Current Accounts

Exports of goods and services, percent of GDP (SeriesExportGoodSer%,
GDP/Capita (PPP) Versus Exports (1995))
Imports of goods and services, percent of GDP (SeriesImportGoodSer%,
GDP/Capita (PPP) Versus Imports (1995))
Aid received, percent of GDP (SeriesAIDRec%GNI, 0.0)
Worker remittances, percent of GDP (SeriesWorkerRemit%GDP, 0.0)
Current account balance, percent of GDP (SeriesXCurActBal%GDP, not used)

The first section of code uses data on current account balances as a check on an estimate of the current account computed from trade, aid, and worker remittances. The reason for this is that data on total exports and imports are not always strong and should be adjusted as appropriate when current account balances are known. This is a procedure that focuses on less developed countries where data are least good.

Total imports and exports (of merchandise and services combined) by country are read as percentages of GDP. Data from the period from 2000 through 2002/2003 (depending on data availability) are averaged to eliminate year-to-year irregularities. Holes are filled using a cross-sectionally estimated function of GDP per capita at PPP; as in many other instances of hole filling, this is unsatisfying but values are required.

Total credits are computed as exports plus aid and worker remittances received and total debits are computed as imports (because of the focus on the poorest countries, aid and remittance payments are ignored). An estimate of the account balance is credits minus debits.

Current account as percent of GDP is read next. Holes are not filled because the only purpose for it at this point is to check on the basic quality of data for its components, especially imports and exports.

When current account data do exist, a difference between the bottom-up estimate and the data can be calculated. When current account data do exist, but there are no data on exports and imports (so that instead, cross-sectionally estimated functions were used above to provide some numbers for those variables), it is likely that there will be a substantial difference between bottom-up estimates and current account data. The difference, whatever its size, is assigned to the imports and exports, directly one-half it to each.

Additional problems will inevitably be created for the model whenever the current account is more than 20 percent of GDP. When the current account is larger than 20% of GDP and either exports or imports are greater than GDP, it can indicate an entrepot economy with what is most likely a transient imbalance in the current account. The transient is eliminated from initial conditions by adjusting the current account down to

20% of GDP; if the gap is positive, exports are reduced and if it is negative, imports are reduced.

After the above potential adjustments, the ground-up calculation for account balance is recomputed. If the imbalance remains more than 20% of GDP, exports and/or imports are recalculated to bring it down to 20%. This is arbitrary, but the presumption is that such imbalances almost invariably represent data problems that should not be allowed to affect the model.

[shouldn't I balance global X and M hear again before decomposing ?–perhaps not necessary, but useful]

5.2.2 Decomposing Trade by Sector: Initial Checks

Now that exports and imports are read and reconciled generally with current account balances, it is time to decompose trade into trade by economic sector, beginning with the use of the physical trade computed in agriculture and energy. Therefore exports and imports of crops, meat, and energy are read from the output files of IFs.mdb to which they were written in their respective preprocessing. Crop and meat are converted to monetary values using crop and meat prices (WORLDAP1 and WORLDAP2) and energy by using energy prices (WEP).

At this point energy production by type of energy is also read from the output files of the energy preprocessor; those values will be needed later for the reconciliation with monetary values in the input-output tables.

The sum of values of agriculture and energy trade are checked against monetary values for total trade to make sure that it is greater than the sum of them and leaves some "head-room" for other merchandise trade and for service trade. If there is not, total exports or imports are adjusted upward accordingly.

A further check is undertaken at this point on total exports and imports. It is to make sure that exports minus imports are less than 80% of GDP, assuring that there is at least some minimal "head room" here for other expenditure components (private and governmental consumption plus investment). Such lack of head-room has appeared in data of major energy exporters such as Azerbaijan, where it is likely that economic statistics have not been adjusted to keep up with rapid growth in oil production and exports. [why isn't this done earlier before moving to physical – were exports or imports recomputed somewhere? Seeing beginning of decomposition below, looks like this should be done after merchandise and service reading.]

At this point exports and imports, maintained until now as percentages of GDP, are multiplied by GDP to obtain absolute values.

5.2.3 Decomposing Trade by Sector: Step by Step

Merchandise exports (SeriesExportsMerchandise, .8*exports)
Merchandise imports (SeriesImportsMerchandise, .8*imports)

Service exports (SeriesExportServices, .2*exports)
Service imports (SeriesImportServices, .2*imports)
Arms imports as percent of total imports (SeriesArmsImp%TotImp, 0.0)

Processing of total exports and imports by country and for the global system is now fundamentally complete. The next series of steps decomposes that total trade into trade by economic sector, maintaining global balances.

The first step in decomposition is to read data on exports and imports of merchandise and services. For the historic load the first year of such data is, unfortunately, 1970 or later, so these values are read. If merchandise trade are missing, a value of 80 percent of total trade is assigned; if service trade are missing, a value of 20 percent of total trade is assigned. [why don't I check to see if other is available and compute as residual percentage?] Merchandise and service trade are then summed and normalized to total trade (exports and imports separately) for each country.

Imports of arms are also read at this time. [not exports? Nothing done with them?]

Service exports and imports are summed globally, and the average value of global exports and imports is used to normalize values and assure that global trade is in balance. Merchandise exports and imports are then recomputed as the residual of total exports and imports at the country level. Then merchandise exports and imports are summed globally and the average is used to normalize country-specific values and assure global balance, while also assuring that merchandise imports and exports are maintained between 20 and 95% of total imports and exports for all countries.

Exports of ICT, percent of exports (SeriesExportICT%Exp,
GDP/Capita (PPP 2000) Versus ICT Expend % of GDP (2001) Exp)
Imports of ICT, percent of imports (SeriesImportICT%Exp,
GDP/Capita (PPP 2000) Versus ICT Expend % of GDP (2001) Exp)

The process of decomposition next moves to ICT trade, which crosses over merchandise and service categories. Data are read for both as a percentage of total exports and imports for countries. Data are relatively sparse and missing data for exports and imports are filled by using a cross-sectionally estimated function of GDP per capita at PPP. Global sums are computed for ICT exports and imports and country values are normalized to the average of the two. At this point ½ of exports and imports by country are assigned arbitrarily to merchandise trade and the other ½ to services – possibly GTAP data would have information that would allow a more informed division.

Next residual service trade is renormalized globally. Residual merchandise trade by country is computed as the total exports or imports minus the residual service trade and minus total ICT trade.

It is possible to move now to agricultural trade. Needed data are available in IFs.mdb, where they were written after processing as described in earlier chapters. Exports and imports are read by country for crops, meat, and fish, with null values (missing data) set to an arbitrary 0.01. World agricultural prices for crops and meat/fish (WORLDAP1 and WORLDAP2) are used to convert these physical values to monetary ones. Global sums are computed for total agricultural trade (crops plus meat/fish) and the average of exports and imports is used to normalized country-specific values. Note: essentially this same process was done earlier in the initial check to make sure that total exports and imports had more than enough room for agricultural and energy trade in value terms.

At this point monetary values of agricultural trade by country are compared with 90 percent of the residual merchandise trade levels and if greater than that value are reduced to it (90 percent is used rather than 100 percent so as to leave head room for energy and manufactures trade). After this adjustment global agricultural trade is again normalized and the residual merchandise trade is computed. An error check is done to assure that all residual values, for both imports and exports, are non-negative.

The above process is then basically repeated for energy trade. Physical energy exports and imports are read and converted with energy price (WorldEP) to monetary values. Those values are bound at no more than 90 percent of remaining merchandise trade and the resultant values are normalized globally. The process of bounding at 90 percent of remaining merchandise trade and normalizing is repeated in case the first normalized pushed any values above that level. Again an error check assures that values are non-negative.

The residual of merchandise trade is recomputed, removing energy trade from both exports and imports.

Exports of ores and metals as percent of merchandise exports
(SeriesOresMetsEx%MerchEx, 0.0)
Imports of ores and metals as percent of merchandise imports
(SeriesOresMetsIm%MerchIm, 0.5)

The process is then again repeated for materials trade. Exports and imports of ores and metals as a percentage of merchandise trade are read (SeriesOresMetsEx%MerchEx and SeriesOresMetsIm%MerchIm), null values are set at 0, and the percentages are multiplied by total merchandise trade to get monetary values. These are again constrained to be less than 90 percent of the residual merchandise trade and are summed and normalized globally. The constraint of 90 percent is reapplied and the normalization is redone. The residual of merchandise trade is further reduced by materials trade and an error check assures that the residuals are positive.

Because the only remaining portion of merchandise trade is trade in manufactures, exports and imports of manufactures are set equal to the residual. Global sums are computed and checked to assure that they are essentially the same.

[Note: we could split this preprocessor return here by either saving and later retrieving the trade values or, probably better, by putting trade into common for the econdata.bas file]

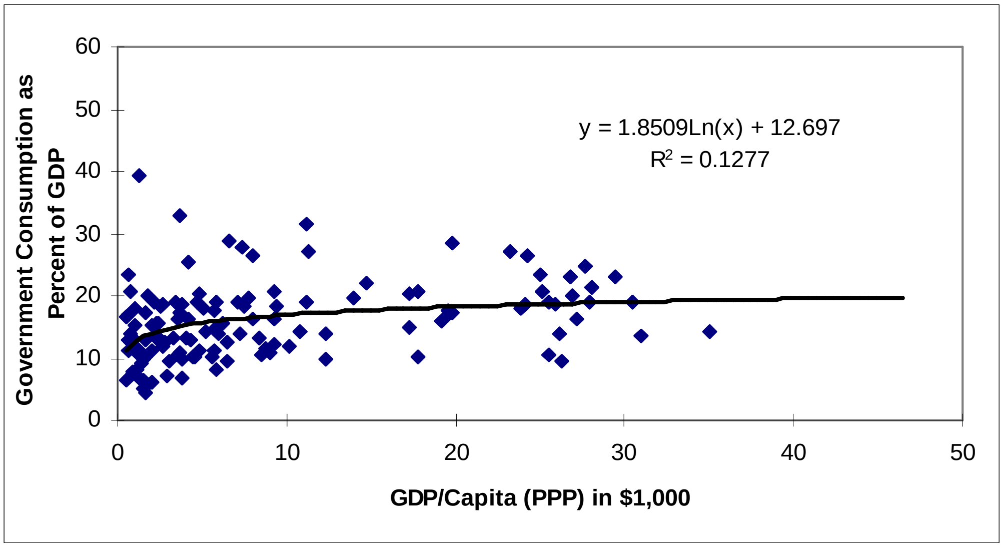
5.2.4 Non-Trade Expenditure Components

Household consumption, percent of GDP (SeriesHouseCon%GDP,
GDP/Capita (PPP) Versus Pers Consumption (1995))

Government consumption, percent of GDP (SeriesGovCon%GDP,
GDP/Capita (PPP) Versus Govt Cons % GDP (2000))
Investment, percent of GDP (SeriesInvest%,
GDP/Capita (PPP) Versus Investment, Fixed (1995))

Now that all global and country-specific trade has been computed, the preprocessor can move to the determination of other expenditures components. Values for private consumption, government consumption, and investment as a percentage of GDP are read for all countries. Null values are filled from cross-sectionally estimated functions of GDP per capita at PPP. Household consumption is constrained to be between 30 and 80 percent of GDP, government consumption between 9 and 35 percent (income transfers can make the total government share of GDP much larger, as we shall see), and investment is constrained to be between 5 and 40 percent of GDP.

The figure below illustratively shows the function estimated cross-sectionally that is used to fill holes in the governmental consumption data.



Total expenditure components (C+G+I+X-M) should be 100 percent of GDP. Household consumption, government consumption, and investment are adjusted as necessary to assure that they are.

5.2.5 Value Added

Agricultural value added, % of GDP (Series, VaddAg%,
GDP/Capita (PPP) Versus Value Added in Ag (1995))
Manufacturing value added, % of GDP (SeriesVaddMan%,
GDP/Capita (PPP) Versus Value Added in Man (1995))
Industrial value added, % of GDP (SeriesVaddInd%,
GDP/Capita (PPP) Versus Value Added in Ind (1995))
Services value added, % of GDP (SeriesVaddSer%,
GDP/Capita (PPP) Versus Value Added in Ser (1995))
ICT value added, % of GDP (SeriesVaddICT%,

ICTShare and algorithm)

Pre-processing next moves to the value-added side of the economy. Value added as a percentage of GDP is read for agriculture, manufactures, industry, services, and ICT. Holes are filled with cross-sectionally estimated functions of GDP per capita at PPP. Some data values for agriculture and services as portions of GDP, notably those reported by countries with GDP per capita at PPP of less than $1000, tend to be unreasonably low or high, respectively. For those countries, values for agriculture are constrained to be no less than 70 percent of those in the cross-sectional function and those for services are constrained to be no more than 30 percent above those of the cross-sectional function.

Value added is needed also for energy and materials sectors of the model. When manufacturing value added is non-null, it is assumed that the sum of value added in those two sectors is the difference between the value added of industry and manufactures (industry is the more general concept), and the difference is split between the two sectors in proportion to their exports. When manufacturing value added is null in the data, exports of manufactures, energy, and materials are used to do a three-way split of industrial value added.

Value added data for ICT are especially sparse. Holes are filled with an exogenously specified ICT share of GDP (ICTShare). The ICT value added is split out from manufactures and services, with the share coming from each determined by their relative size but not to exceed 90 percent of each.

To assure that neither data nor adjustment processes have dramatically unbalanced the sizes of manufacture and service value added, the former is constrained to be at least 5% of the sum of the two.

The sum of value added across all sectors must equal GDP and that is assured via a final normalization in this part of the preprocessing.

At this point, the fundamental economic elements are structured for each economy of the world and the trade balances are assured. It is time to move to the intersectoral flows and the broader social accounting matrix balances.

5.3 IO Matrices and Social Accounting Matrices

5.3.1 Reading and Aggregating Intersectoral and Value Added Data

The first step in the process of moving to intersectoral flows is the computation of input-output matrices for each country in the 6-by-6 sectoral format of IFs. Raw data come from the Global Trade Analysis Project (GTAP). Its 60+ sector matrices are collapsed to 6 sectors using a concordance table that is stored in the GTAPSectorConversion table of IFs.mdb. Although the GTAP project covers a large number of countries, its matrices do not cover the full 182-countries of IFs. It does, however, provide regional values that can be used to flesh out all 182. Nor are its matrices set up for dynamic forecasting; as a result of economic growth and technological change, matrices for individual countries should change over time. In order to facilitate forecasting, IFs creates a set of generic IO matrices at different levels of development (as represented by GDP per capita at PPP). All of these steps are undertaken in a subroutine of the economic preprocessor called InputOutputData.

The concordance table is read first. Then, for all of the IFs countries, flow data are read for the entire GTAP input-output matrix as well as for value added of land, unskilled labor, skilled labor, and capital. The "regional" matrices and value added vectors of GTAP (for instance, that for sub-Saharan Africa) are used, as appropriate (correspondences are stored in a field called GTAP Country Codes) to assign GTAP values to each of the IFs countries. The GTAP dataset distinguishes between intersectoral flows from domestic sources (Flows.csv) and those from imports (ImptFlows.csv). Both need to be read, and in the computation of the technological matrices the two need to be summed.

The next step is the aggregation of the full GTAP matrices and value-added vectors into the 6 sectors of IFs using the concordance table (cells of the larger matrix are simply collapsed and summed into cells of the smaller matrix). Then the columns of the matrices plus the appropriate sectoral value addeds are summed and the sum is used to compute the input-output matrices; by definition the IO coefficients are the intersectoral flows divided by the column sum. A check is done to make sure that the row sums of the resultant matrices are less than 0.95 and, if not, the cells are proportionately reduced (there must be a type of head room here also, leaving space for deliveries to expenditure components as well as to other sectors). The resultant values are saved into the IFs.mdb file for later use.

5.3.2 Generic Matrices for Dynamic Use

The flow moves at this point to the computation of the generic input-output matrices for dynamic use. IO coefficients for each of the 182 countries are summed in one of x categories by level of GDP per capita (where x is currently less than 10) and averaged across the countries added to the category. X is currently 8 and the breakpoints are $100, $250, $500, $1000, $2000, $4000, $6000, $8000. The same is done for value added of land, capital, and the two labor types. The resultant generic matrices are again saved to tables in the IFs.mdb file.

5.3.3 Expenditure Components

Although value addeds have been specified by sector in the portions of the preprocessor described above, the expenditure components have not. Government spending by origin sector is assumed to be almost entirely from the service sector (90 percent) with only 10 percent coming from manufactures. Investment is assumed to come from manufactures and services in equal shares.

[Note: because there are value added data in the GTAP matrices, these could be summed into sectors and used for the countries in the GTAP set instead of the procedure above that relied on partial data (for instance, no breakdown of energy, materials). That would seem likely also to give a value added more consistent with the IO matrices for those countries and simplify the adjustments now made to assure such consistency]

The expenditure component that is particularly important to specify with greater accuracy, and to prepare for dynamic forecasting is the largest of them, household consumption. Given the input-output matrices, the value added by sector, sector-specific trade and crude estimates of government and investment by origin, it can be computed.

Within IFs, total household consumption is spread across five sectors: agriculture, energy, manufactures, services and ICT. It is assumed that households do not directly consume other primary raw materials. It would theoretically be possible to take household consumption by sector of origin from survey data rather than computing it. In general, however, value added data are stronger than household consumption data, and we privilege them accordingly. Doing so has the added, and large benefit, of assuring that national accounts will balance across value added, intersectoral flows, and final demand.

Household consumption in IFs is assigned initially to three sectors: agriculture, manufactures, and services.   That is done by summing across the input-output matrix row for the respective sector and dividing the value added of the sector by 1 minus the row sum in order to compute the gross production of the sector. Gross production minus the sum of the actual flows within the row provides a calculation of the amount of the sector’s production that is delivered to final demand. That estimate, augmented by imports and reduced by deliveries to other sectors of final demand including exports, provides an estimate of deliveries to households. (Arms imports, which for some countries like Eritrea can be very high in the raw data, are explicitly removed from manufactures in this calculation so as not to artificially augment household consumption of them in the computation).

The consumption in the three sectors is computed in the above procedure as portions of total household consumption. Energy consumption is set at 5 percent of total household consumption, a rough global average.   There is, however, also information about energy consumption that comes from the physical energy data. Apparent energy consumption was computed earlier in processing of energy data (as production plus imports minus exports). Seldom is more than 60 percent of energy consumption consumed by final demand (the rest satisfies intersectoral flows). Thus 60 percent of apparent consumption in physical terms, times the world energy price, is used as an upper limit of final energy demand. If the 5 percent specification times total household consumption exceeds that maximum, the share of household consumption assigned to energy is reduced accordingly.

[Why don't I handle ICT value added like ag, man, serv? value added in ICT was already split up with ICTSHARE and have IO coefficients collapsed appropriately. Probably procedure below is pre-use of value added for consumption and needs revision]

The other complicated sector of final demand to assign is ICT. ICT, as elsewhere in IFs, is assumed to be imbedded in the manufacturing and services. The calculation of personal consumption of ICT builds from the exogenous ICTShare specification. The absolute amount of household consumption of ICT is the ICTShare of the percentage shares of manufactures and services, augmented by ½ because manufactures and services are the largest portions but not the total of household consumption. [I don't like this – revisit; probably because only applied to man and ser rather than total C; could compute man and ser percentage shares and enhance accordingly, but entire procedure for ICT might be better done from value added, as suggested above]. The manufactures and service sector shares are reduced accordingly.

The final step is to sum all of the shares of household consumption and to normalize them to 100 percent of the total, simultaneously converting from percentage shares to values.

[Note: try replacing this ICT calculation with exactly the same procedure as for first three sectors; save base, rebuild, run a couple of years, and check ICT share in both approaches]

5.4 First Steps on the Social Accounting Matrix

At this point the basic economic elements, value added, intersectoral flows, and final demand, have been computed. It is time to begin moving into the elements of the social accounting matrix. For a general understanding of the approach of IFs towards building and using universal (all countries), integrated (balanced across countries) social accounting matrices see Hughes with Hossain (2003).

5.4.1 SAM: Government Revenue and Expenditure Flows

The matrices are built up in a series of steps. The first steps focus on government consumption and government revenues, in part because these data tend to be stronger than many others in the matrices. Specifically, the following variables are read:

Corporate taxes (SeriesTaxCorp%Tot, GDP/Capita (PPP) Versus
Corporate Tax % of Total (2000))
Welfare/social security taxes percentage of revenue (SeriesTaxSocSec%CurRev,
GDP/Capita (PPP) Versus SS Tax % of Total (2000))
Taxes on goods and services as percentage of revenue (SeriesTaxGoodSer%Cur Rev,
GDP/Capita (PPP) Versus Indirect Tax % of Total (2000))
Government revenue as a percentage of GDP (SeriesGovtCurRev%GDP,

GDP/Capita (PPP) Versus Govt Revenue (2000))
Government expenditures as a percentage of GDP (SeriesGovtExpend%GDP, GDP/Capita (PPP) Versus Govt Exp % GDP (2000))
Government pension spending as a percentage of GDP (SeriesGovtPension%GDP, GDP/Capita (PPP) Versus Public Pension % of GDP (2000))

Values for each of the above are filled from cross-sectionally estimated functions when data are missing. Government revenue is bounded to be at least 5 percent of GDP.

Household tax shares of total revenues are computed as the residual of 100 percent minus the corporate/firm, indirect, and welfare/social security taxes. Household taxes are assumed to be at least 1 percent of the total revenues and other tax shares are reduced accordingly if they are not. Note: this conceptualization leaves out non-tax revenue such as that which might be obtained from government-owned industry.

At this stage, identical household tax rates are being assigned to unskilled and skilled households (which implies no progressivity or regressivity in the tax rates computed as a share of income); we are looking for data on relative tax burdens, but absence of differentiation is not a bad initial working assumption.

$$HHTaxShr_{r,h} = 100 - IndirectTaxShr_r - FirmTaxShr_r - SSWelTaxShr_r$$

Given shares of total tax revenue provided by firms and households, as well as collections for social security/welfare, the pre-processor can compute tax rates on income as the tax collections divided by household and firm incomes. A necessary prior step is to specify household and firm incomes. IFs uses a cross-sectional function of GDP per capita to compute the value added share of firms, and using a Cobb-Douglas approach to GDP assigns the remaining share to households: GDP/Capita (PPP) Versus Cobb-Douglas Alpha (GTAP 5). Both shares are multiplied by GDP to get total firm and household income values. [Note: now that have GTAP shares, could/should use those and fill holes/shift perhaps with function]

The computed rates are those used above in the annual calculations of the model.

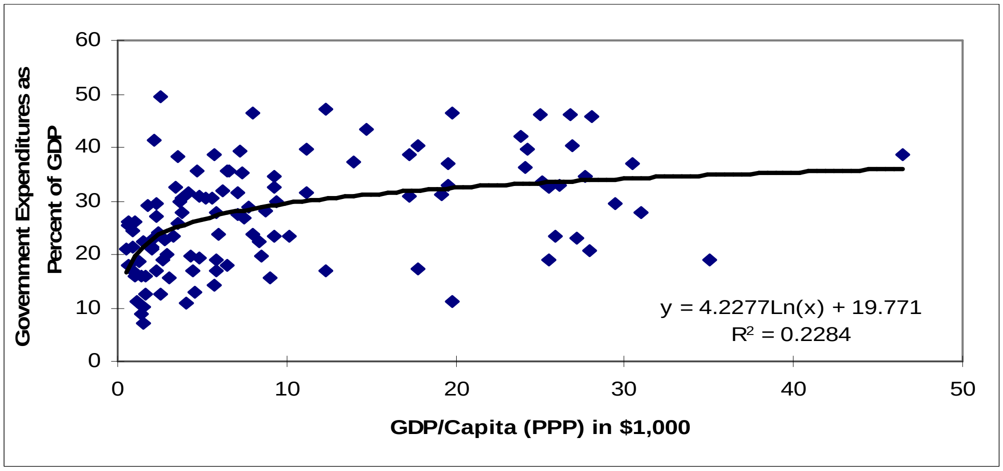
$$FIRMTAXR_r = GOVREV_r^{t=1} * FirmTaxShr_r^{t=1} / FirmInc_r^{t=1}$$
$$HHTaxR_{r,h} = GOVREV_r^{t=1} * HHTaxShr_r^{t=1} / HHInc_r^{t=1}$$
$$SSWELTAXR_r = GOVREV_r^{t=1} * SSWelTaxShr_r^{t=1} / (FirmInc_r^{t=1} + HHInc_r^{t=1})$$

[Note: indirect tax equation INDIRECTTAXR was left out of the above set in Economy and SAM document]

IndirectTaxR = GOVREV * IndirectTaxShr/FirmInc

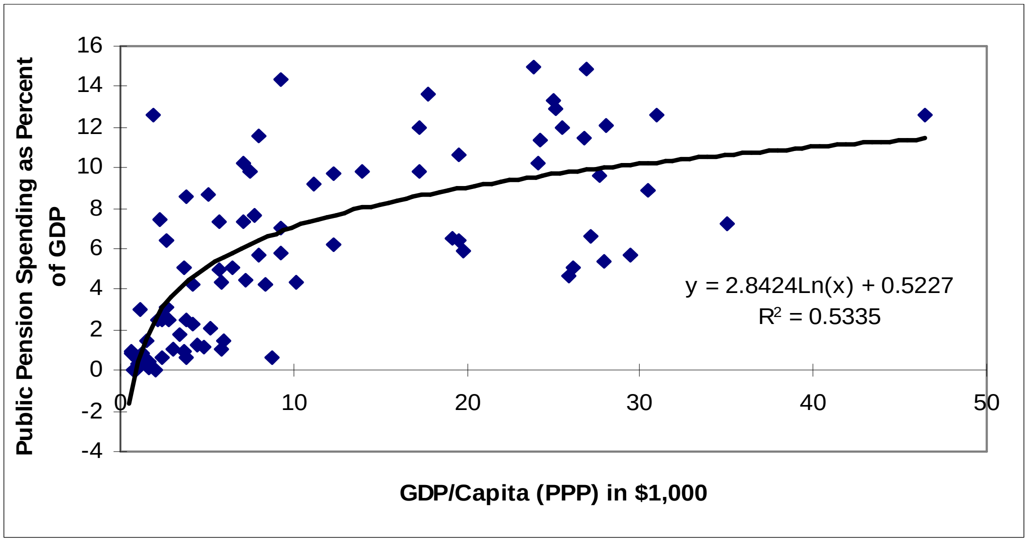
Government expenditure data are not as good as revenue data and the cross-sectional estimation of it with GDP per capita is not very strong. Government consumption data are better, however, so if expenditure data (including transfers) do not exist, they are tied to consumption.

After reading government expenditures and filling missing values from a cross-sectionally estimated function, foreign aid receipts are added. For some large aid recipients (like the Democratic Republic of the Congo), it is impossible to work on governmental accounts without accounting for receipts – in contrast, donations are nearly insignificant in the accounting of donors. The function for filling holes, using data from the WDI 2002, is shown below. It is GDP/Capita (PPP) Versus Govt Exp % GDP (2000).



A check is made to make sure that total governmental expenditures are at least 1 percent greater than governmental consumption (read and used earlier). Normally, transfer payments will make it considerably higher.

Government expenditures consist of a combination of direct consumption/expenditure and transfer payments. As a general rule, transfer payments grow with GDP per capita more rapidly than does consumption. And within transfer payments, pension payments are growing especially rapidly in many countries, especially more-economically developed ones. The figure below shows the relationship between GDP/capita and public spending on pensions as a percent of GDP (using data from a World Bank analysis of data drawn from OECD and other sources). This function is used to fill holes in the data set. It is GDP/Capita (PPP) Versus Public Pension % of GDP (2000).



5.4.2 SAM: Public Finance Stocks

At this point, the preprocessor moves to a computation of some basic Social Accounting Matrix stocks. It does so because some are useful for analysis of subsequent flows.

The first steps in looking at stocks that surround public finance are to read World Bank loans and IMF credits:

World Bank loans as percent of GDP (SeriesXWBLoans%GDP, 0.0)
IMF credits as percent of GDP (SeriesXIMFCredit%GDP, 0.0)

When these series are read, missing data is assumed to be zero debt or credit. Percentage values are multiplied by GDP to get the actual size of debt to the World Bank and IMF. Global totals are computed of debt to both institutions.

For initial stocks to balance globally, there normally should be off-setting credits for these debits. There is no database for the subscribers to the two international financial institutions (IFIs). IFs arbitrarily allocates the credit side to the economically-advanced country proportional to the size of their economies. Strictly speaking, this step is probably unnecessary, because it is arbitrary and such countries never expect to receive official state funds back. Yet many IFI fundings are actually private sector-based, and there is debt to be repaid and a reverse flow that will be generated.

The logic turns next to more general public and private debt stocks. The preprocessor reads total public or publicly-guaranteed debt and private non-guaranteed debt:

Public and publicly-guaranteed debt (SeriesXDEBTPPG%GDP,0)
Private non-guaranteed debt (SeriesXDEBTPNG%GDP,0)

All of the private debt and one-half of the public and publicly-guaranteed debt is allocated to debt of firms (XFIRMDEBT). [Probably this should vary by development level, with more debt for the poorest allocated to government].

Unfortunately, we have no data for the source of origin of this indebtedness, but for a balance of stocks need to allocate it to more developed countries. This is done proportionally to GDP for countries with more than $10,000 per capita. [Could consider allocation tied to integration of current account information for the wealthy as total debt, below]

The next step is the preparation of total international debt positions, using data when available and integrating the current account over a long period when it is not. As foundation values are read for

> Exports of goods and services, percent of GDP (SeriesExportGoodSer%, 0.0)
> Imports of goods and services, percent of GDP (SeriesImportGoodSer%, 0.0)
> Aid donations, percent of GDP (SeriesAidDon%GNI, 0.0)
> Aid receipts, percent of GDP (SeriesAidRec%GNI, 0.0)
> Aid receipts, percent as grants (SeriesAIDRecGrant%Total, algorithmically 0.0)
> External debt (SeriesXDEBT, accumulation algorithm)

An initial calculation of total accumulated external debt (XDEBTACC) is made by integrating the trade over as much of the 1960-2000 period as data are available and adding the net flows of foreign aid (the non-grant or loan share only).

Then actual debt/net external position data are read when available, which is fundamentally only for developing world debtors. When a country is not shown in the data as a debtor, it is assigned the accumulated net external position from the integration process (and may become a creditor or debtor as a result).

When a country is a debtor in those data, a check is made to assure that the total debt level is equal to or greater than the sum of the indebtedness to IFIs and the foreign indebtedness of firms (both calculated earlier). If a country is a creditor, a check is made to assure that the positive external position is larger than the contributions to IFIs and the loans to firms abroad, both also calculated earlier.

The calculations of external debt, relying heavily on data, are assumed to be of higher quality than the calculations of external assets, relying mostly on the integration process. Thus instead of the typical normalization to the average of global totals, a normalization is made of creditors to the global total of indebtedness.

The logic next moves to the calculation of governmental external and domestic indebtedness. A calculation of total government external debt (XGOVTDEBT) is made as the difference between total external debt and external firm debt (households are implicitly assumed not to carry external debt). A calculation is made of what is termed “basic” governmental debt as the difference between total external governmental debt and that owed to the World Bank and the IMF. That is, "basic" governmental debt is owed to other governments or, less likely, to firms (banks).

5.4.3 SAM: Back to Flows for Foreign Aid

Next the flow turns to foreign aid, with donations, receipts, and grant shares read once again, this time to calculate initial conditions for the model rather than for the purposes of integrating over time:

> Aid donations, percent of GDP (SeriesAidDon%GNI, 0.0)
> Aid receipts, percent of GDP (SeriesAidRec%GNI, 0.0)
> Aid receipts, percent as grants (SeriesAIDRecGrant%Total, algorithmically 0.0)

Data for foreign aid exist beyond the model base year of 2000. Aid receipts as a percentage of GDP can vary quite widely, so the years 2000-2003 are used to compute an average value. That value is bound at 50 percent for purposes of initializing the model; although aid inflows higher than that can occur in emergency situations (such as the Democratic Republic of the Congo) they are not sustainable in forecasting and would very quickly result in a bad case of Dutch disease. Normalization of global aid donations and receipts to the average of the two completes the process.

5.4.4 SAM: Back to Stocks for Government Debt, Reserves, Financial Assets

Attention moves to additional stock values for the social accounting framework. The first is total government debt as a portion of GDP (the earlier attention was only to the international portion of debt). When data holes exist, they are filled with a cross-sectionally estimated function.

> Government debt, percent of GDP, (SeriesGovtDebt%GDP,
> "GDP/Capita (PPP) Versus Govt Debt % GDP 2000")

Given total government debt and the earlier computation of external government debt, domestic government debt (GOVTDEBTD) is calculated as the residual.

International reserves are read when available and set arbitrarily at 15 percent of GDP when not.

> Foreign reserves, percent of GDP (SeriesXReserves%GDP, 15)

The next stock variables read are FDI inflows and outflows, and portfolio investments (bonds and equity). In each case null values are assumed to be zero. The series are integrated over as many years as possible since 1960 to obtain initial conditions for stock levels.

Foreign direct investment inflows, percent of GDP (SeriesXFDIInflows%GDP, 0)
Foreign direct investment outflows, percent of GDP (SeriesXFDIOutflows%GDP, 0)
Foreign bond purchase inflows, percent of GDP (Series XPortBonds%GDP, 0)
Foreign equity purchase inflows, percent of GDP (SeriesXPortEquit%GDP, 0)

5.4.5 SAM: Back to Flows, International Financial Exchanges

Attention then shifts again to flows. The above FDI and portfolio inflows and outflows are again read, but with the intent of calculating initial rates of flow as a percentage of GDP. Because flow levels can vary substantially year to year, a range of recent years is used to compute an average flow rate (5 years for the 2000 database load and 10 years for the historic 1960 load).

Moving to flows from and to the IFIs, the data only has flows from them.

Inflows from the IBRD, percent of GDP (SeriesXFlowsIBRD%GDP, 0)
Inflows from the IDA, percent of GDP (SeriesXFlowsIDA%GDP, 0)
Concessional inflows from the IMF, percent of GDP (SeriesXFlowsIMFCon%GDP, 0)
Non-concessional inflows from the IMF, percent of GDP (SeriesXFlowsIMFNonCon%GDP, 0)

Therefore the procedure is to initial the inflows from the IFIs with data, averaged over 5 recent years, and to initialize the outflows back to the IFIs based upon the debt to them calculated earlier. At this time, flows from the World Bank sum those from the IBRD and IDA and flows from the IMF sum concessional and nonconcessional. For the calculation of flows back, interest rates (xwblintr, ximfcrintr) are applied to the total debt levels.

Moving to worker remittances, once again the data show the receipt of them as a percentage of GDP, but not the states of origin. In order to assign remittances back to states, information on the foreign population percentage is used.

Worker remittance receipts, percent of GDP (SeriesXWorkerRemit%GDP, 0)
Foreign-born population, percent of total (SeriesPopForeign%, 0)

In some cases, poor and remittance-receiving states can also have foreign workers (e.g. a Djibouti with refugees from other African states). These countries are assumed not also to be remittance sending. When states have GDP per capita of at least $3,000, have foreign populations, and are not shown in the data as remittance receiving, they are placed into the remittance sending pool. In subsequent code they will be assigned a portion of the global remittance pool.

At this point a number of calculations are actually made, the ground for which was prepared earlier. These are primarily to normalize global flows and to assign flows back to sources of origin. Specifically, they are:

> Normalize flows into states from the World Bank (XWBLNFIN) to the global total of new outflows from the Bank and repayments of debt to it (assuming that debt repayments are re-circulated).
>
> Normalize flows into states from the IMF (SIMFCRFIN) in the same way.
>
> Allocate the stock of outward portfolio investment back to developed countries (defined by GDP per capita greater than $10,000) by economic size, in order to assure that the total stock of inward portfolio investment is matched globally by the stock of outward investment. Then the current outward flows are computed so that the total matches total inward flows, allocating the outward flows proportionately to the stocks of outward investment.
>
> Outward FDI stocks were computed earlier based on integration of flows. But the outward data are not as good as the inward data. So FDI outward stocks are normalized globally to the global total of inward stocks.
>
> The actual assignment of outward worker remittances is done, based upon the preparation done earlier for assignment to countries with GDP per capita above $3,000. Again the global normalization is to the world sum of inward flows.

5.4.6 SAM: Attention to Problem Countries

The next and essentially last step in the preparation of the data for the social accounting matrices, both flows and stocks, is attention to the states that have substantial discrepancies in their data for financial inflows and outflows as indicated by the figures reported for current and capital accounts relative to debt and reserve situations. In some cases this could simply indicate errors in the data available, for instance unreported/uncollected data on worker remittances. In others it could suggest capital flight or theft of monies generated by state-controlled natural resource production. The intent here is to calculate a variable called unexplained finances (XFINUNEXP) for display and possible future use in the model.

For help in these computations a number of variables are read from data files:

> Public and publicly-guaranteed debt (SeriesXDEBTPPG%GDP,0)
> International reserve holdings (SeriesXReserves%GDP,0)
> GDP in 2000 dollars (SeriesGDP2000)
> Financial payments abroad (SeriesXIncPayments%GDP,0)
> Financial receipts from abroad (SeriesXIncReceipts%GDP,0)

The procedure involves comparing the change in international position of countries with respect to debt and reserve holdings (IntlPositionChange is reserve growth minus debt growth) with a computation of expected change in that position based on current and capital account flows (CompIntlPositionChange). The calculation is a "rough and ready" effort to highlight the countries with most significant problems in such data (like Angola, but also Russia).

The change in actual international position is computed from the change of public and publicly-guaranteed debt and the change in reserve holdings. Because of weaknesses in the debt and reserve data, the two most recent years are compared with the two prior years and the change in position is estimated at half of the difference. The change in international position is the growth of reserves minus the growth in debt.

The computed or apparent change in international position sums current and capital account terms. The current account terms are exports minus imports plus aid received minus that given, plus net worker remittances, minus interest on debt to international financial institutions, external firm debt, external government debt, and portfolio holdings by foreigners. The capital account terms are net FDI inflows, net portfolio inflows, and net inflows from IFIs. Unexplained finances are the actual change in international position minus the computed change, as a percent of GDP. They are saved and placed into a model variable named XFINUNEXP so that the user can see this problem cases.

There are, however, some special problems surrounding data on worker remittances. In particular, Lesotho has workers in South Africa that provide a significant financial input to the economy but who do not show up as migrants and in the remittance data. Therefore a second check on international positions is made. Actual net income from abroad according to data are financial receipts minus payments. Apparent or computed net income is the sum of worker remittances plus the interest term on all foreign indebtedness noted above with respect to the current account. A possible remittance discrepancy term is computed as the difference of the actual and computed net income terms. If the term is positive and a country is a net recipient of remittances, those remittances are increased accordingly, up to a maximum of $5000 per worker abroad. A downward adjustment to the unexplained financial term is made accordingly.

It also appears that debt forgiveness in the Heavily Indebted Poor Countries initiative may not be captured in the current data series. At this point no correction of this omission has been made.

5.5 Initialization of GDP Growth and Income Distribution

The processes described above complete the initialization of the social accounting matrices. The flow shifts next to the computation of initial values for one of the most important variables in the model, GDP growth rate. It also reads and initializes income distribution in terms of the Gini index:

GDP in 2000 dollars at MER (SeriesGDP2000, required)
GDP in 2003 dollars at PPP (SeriesGDP2003PPP, not used when null)

GDP growth varies widely from year to year. Setting the initial value for GDP growth rate in the model (IGDPR) is very important because it determines both the initial calculation of multifactor productivity (MFP) and, of course, the initial pattern of growth in each country.

The preprocessor looks to the years from 1990 through 2003 for the basis of computation of long-term economic growth rates, using data in 2000 dollars at market exchange rates. It weights the first six years of that period, through 1996, half as heavily as the more recent 7 years. Compound annual growth rates are computed for each of the two sub-periods. For those very few countries without GDP data, a growth rate is estimated from a cross-sectionally estimated function (GDP/Capita (PPP) Versus Economic Growth Rate (1995)). The final value for IGDPR is bound to be at least 0.5 percent higher than population growth and to be no more than 8 percent. For the historic load values from 1960 through 1980 are used to set initial values.

At this point initial interest rates (INTR) are set at the initial GDP growth rate, but are bound between 2 and 8 percent.

In addition, economic growth rates are computed for the years 2000, 2001, 2002, and 2003 using data from the GDP file in 2003 at purchasing power parity. These are saved in a variable (GDPR) that is used to override actual model computations of this important variable in the first three years of the model's execution, and the series will be extended over time.

[Now that have MER data past 2000, should switch over to it from the PPP series.]

This section of the preprocessor also computes some important income distribution variables using three data series:

GINI indices, constructed series (SeriesGINIExtended,
GDP/Capita (PPP) Versus Gini Index (2000))
Percent living on less than $1 (SeriesIncBelow1Dollar%,
GDP/Capita (PPP) Versus Income Below One Dollar per Day (2000))
Percent living on less than $2 (SeriesIncBelow2Dollar%,
GDP/Capita (PPP) Versus Income Below Two Dollars per Day (2000))

The numbers are read and null values are filled using the indicated cross-sectionally estimated function.

This completes the bulk of the preprocessor's basic reading and filling of data values. At this point it writes all of those computed to the EcoOutput table of the IFs.mdb file.

5.6 Adjusting Input Output Data

At this point the preprocessor has handled input-output data from GTAP, during which it computed initial IO tables for each country and generic tables for different development levels. It has also processed all initial economic data, filling holes, creating social accounting matrices for both flows and stocks, and reconciling many disparate data sources.

It has not, however, reconciled the input-output and the broader economic data. The flow thus turns to doing so, calling two subroutines:

```
Call InputOutputDataAdjuster
Call RegionalInputOutputData
```

The InputOutputDataAdjuster routine begins simply by reading back into memory a number of key elements: the generic IO matrices; associated GDP per capita thresholds; the country-specific IO matrices computed earlier.

5.6.1 Reconciling Input-Output, Value Added and Final Demand Data

The first important reconciliation undertaken is of sector-specific production for final demand and sector-specific value added to each other, taking into account the values of the IO matrices. Value added for all sectors was computed earlier (although data really only exist for agriculture, manufacturing, and services). [Note again: should look to GTAP data for improvement in Value Added computations earlier.] Production for final demand (PFD) is a sum of sector-specific household consumption, government consumption, investment by origin sector, and exports, minus imports.

By definition the sum of value added and the sum of this production for final demand both equal GDP, and the processes to this point have assured that they will do so. Most of the value-added specifications can be assumed to be reasonably good, and the specifications of final consumption for agriculture and energy are tied to the physical values as discussed in earlier sections, also giving them reasonably sound values. There is no direct use of other raw materials for household or government consumption or for investment by origin. Therefore only the trade values enter production for final demand and those are reasonably good. There is limited magnitude of ICT final demand.

The key values to reconcile are therefore in the manufacturing and services sector. If neither value added nor final demand values are to be altered in other sectors, and given that the sum of all value added and all production for final demand equal GDP, the reconciliation involves shifting values for final demand between manufacturing and services.

In addition, the value added specifications for those sectors are reasonably sound. It is the allocation of household consumption to them, as well as the division of government consumption and investment by origin between them that is understood to have earlier been quite arbitrary – there are not data for these divisions across countries. Therefore this reconciliation process focuses on adjustments of final demand allocations between manufactures and services, assuming that the value added and IO data should be privileged.

The beginning step in the reconciliation is to compute gross production by sector from the final demand side and from the value added side. From the final demand side, the process uses an inversion of the input-output matrix (subroutine EInvert, not documented here but available in the code), the input for which is the PFD vector and the output of which is gross production by sector (ZSfromPFD). ZSfromPFD is summed for manufacturing and services, and the manufacturing share is computed: ZSManShrFromPFD.

Turning to the value added side, the individual IO coefficients are summed down each column, and the gross production by sector (ZSfromVadd) is computed as the respective value added divided by 1 minus the column sum of the IO coefficients. ZSfromVadd is summed for manufacturing and services, and the manufacturing share is computed: ZSManShrFromVadd.

The discrepancy in the two calculations of gross production is computed as the PFDVADDGap. An iterative process is then undertaken to correct portions of that gap by shifting some of the household consumption, government consumption, or investment by origin from manufacturing to services or vice versa. Small portions are shifted, the production for final demand is recomputed, the inversion is redone to calculate gross production based on final demand (ZSfromPFD), the manufacturing share is recomputed and compared again with the manufacturing share from the value added side. The shifting of final demand continues until convergence (or 1000 steps towards it).

5.6.2 Reconciling Physical Energy Data with Intersectoral and Final Demand Data

The second important reconciliation turns to the energy sector. Here there are no value added data, so we feel free to recompute earlier estimates as appropriate. And, as noted earlier, the final demand data, linked to the physical side, are reasonably good. So here the primary concern is that physical data on energy production, converted to monetary values, may not be consistent with the gross production of energy computed when the IO matrix was inverted and used to compute it from the final demand side. Given that IO data for many countries have been estimated from regional matrices of the GTAP project and that IO values in that project are not always reconciled with physical data, there can be substantial inconsistencies around energy. The logic of the reconciliation privileges the physical data for total production and trade, and adjusts the IO coefficients and, by default, the value added values, accordingly.

The beginning step of this reconciliation is the conversion of energy production data, summed across all forms, to value terms (TotEnProd). This conversion uses the price of oil, an obvious oversimplification, but not a brutal one. For later use apparent energy consumption (AppDemand) is also computed as total production plus imports and minus imports, all in value terms.

The ratio of TotEnProd and the energy column calculation of ZSfromPFD should be one. An iterative process is begun that moves the ratio towards 1.0. If it is not within the range of 0.99 and 1.01, a multiplier is calculated for application to the energy column coefficients of the IO matrix. Using that multiplier, the IO matrix is again inverted with production for final demand (PFD) as an input and ZS (from the PFD side) as an output. This process is continued with increasing or decreasing values of the multiplier until convergence within the limits or 2000 cycles. An error message is given to the user if the convergence has not yet fallen within at least the range of 0.25 and 5.0. Within that range the forecasting equations can still handle discrepancy, but outside of it the model performance with respect to the linking of energy and the economy will be very weak. The major problem cases are almost invariably producers for which energy is a very large and recently changing portion of the economy (like Azerbaijan).

At the end of this iterative process, the IO coefficients for the country are changed to take into account the multiplier. A check is then undertaken to make sure that the column sums of IO coefficients have not been adjusted upward above 0.95. If they were, that would squeeze value added out of the system and a reduction is made accordingly.

The next check is across the rows rather than down the columns of the adjusted IO matrix. The energy demand of intersectoral flows is computed as the sum of the coefficients for the energy row times the gross production (ZS from the PFD side). That number should not be greater than the total apparent energy demand (AppDemand) computed earlier; if it is, there is no room for final energy demand. In fact, the forecasting model can again handle considerable discrepancies in its own adjustment processes. If the ratio of intersectoral flows to apparent demand exceeds 5.0, an error flag is given to the user during rebuild of the base case.

5.6.3 Moving Towards Closure: Repetition to Narrow Descrepancies

At this point the adjusted IO coefficients are written into the IFs.mdb file, the IFsIOCoefCalcCountryEnAdj table and copied into the matrix AMatCountry for further use by the preprocessor.

Although the sequence of adjustments have now completed any to be undertaken on the IO matrices, some additional checks and some repetitions of earlier ones should now be undertaken. The first involves value added. PFD was shifted as necessary to be consistent with the IO matrices and value added, followed by recomputation of the IO matrices themselves to be consistent with physical energy data. At this point the flow moves to recalculation of valued added, using gross production by sector and subtracting the intersectoral flows down the columns. The newly computed value addeds are normalized to GDP.

[Note: could this be where the SAM totals by column in the sectors somehow become different from the SAM totals by row in the sectors (with the sums by row and column across the sectors still producing identical totals?]

Because the IO coefficients have potentially been adjusted for at least a small number of countries, it is necessary to recompute gross production and to again assure consistency of production for final demand with it and value added. This is done with exactly the same procedure described earlier. Using production for final demand (PFD), a matrix inversion operation computes gross production from the PFD side (ZS from PFD); it is compared with ZS from value added, and adjustments are iteratively made to household consumption, government by sector of origin, and investment by sector of origin. [Note: this procedure, used in more than one place, should be put into a function and removed from the main routine.]

The adjusted gross production (ZS from PFD) is then used with the IO matrix once again to calculate value added for all sectors (VADD) and the resultant figures are normalized to the GDP.

The final step in this process is once again calculating gross production based on value added and the IO matrix this time (ZS from VADD). The reason for this is that the actual economic forecasting in IFs is ultimately based on calculating valued added first (in a Cobb-Douglas formulation), gross production based on value added next, and availability of production to meet final demand as the third step. Values for gross production (ZS from value added), production for final demand, and the sectors of household consumption, government consumption, and investment are saved to the EcoOutput table of the IFs.mdb file.

The above process of adjustments of IO matrices with valued added and production for final demand is both complex and ultimately not completely successful for all countries. The energy forecasting model itself needs to use multipliers computed in the first year to reconcile numbers from the value-based calculations of the economic model with the physical numbers. Those multipliers persist over time and clearly indicate continue discrepancies in the overall process. The aim of the reconciliation process is to reduce those persistent discrepancies to levels that do not produce unacceptable model behavior.

It in essence the process described above involves an iteration (adjusting final demand sources) within an iteration (connecting physical energy production with value-based energy production). The exterior iteration could be repeated more than the two times currently done and would likely converge further. But the foundational raw data available are sufficiently weak and that is not easily overcome. The two steps move far enough towards usability so that the process functions adequately.

The final subroutine accessed in the preprocessor of economic data is RegionalIOData. In the full-country version of the model most often used in recent years, this has no real purpose. Yet it is maintained in case users do wish to reduce the number of regions from the number of countries (currently 182) via aggregation. When countries are combined into regions, the routine averages the IO coefficients with weighting by the GDPs of the countries involved.

6. Education

The education module calculates the student and budgetary flows at three different levels of education, primary, secondary and tertiary. It also calculates the stock of human capital in the population and distinguishes the stock according to their levels of education. Other than the budgetary variables, all the flows and stocks are gender disaggregated. The demographics required for this module are supplied by the population module of IFs. Total educational budgets for countries follow from the government expenditures calculated in the IFs economic module. On the other hand, educational outcomes drive economic growth and some demographic variables.

The preprocessor of the education module (DataEducate.bas), fills up the initial value of the flow rates either from empirical data for the appropriate or nearest year or by using estimation techniques like cross-sectional function or trend extrapolation. Because of their unavailability during preprocessing, detail age-cohort population structure data could not be used in education pre-processor for purposes like consistency check [this is for record only, mti].

6.1 Three levels of education: Student Flows

In IFs the educational life of a student is divided into three successive stages- primary, secondary and tertiary. The dominant dynamics of the student flow at each level starts with an intake, who flow through the grades or dropout annually, until they graduate from the level. The major variables are, generally, intake rate, survival rate to the last grade (used to calculate dropout rate), enrollment rate and graduation rate. We have all the time-series required to initialize the grade-wise student flows in Primary. Secondary and tertiary data, on the other hand, do not fulfill that minimal requirement and we have to make some assumptions to overcome those limitations.

Sometimes, e.g. for budgetary purpose, gross rates are required in addition to the net rates. In all cases of student flows, the rates are calculated for male, female and total (i.e., both genders). The length (duration) of primary and secondary education differ by country. Duration data are obtained from the UNESCO Global Education Digest 2005 and attached to the countries in the pre-processor. These durations are maintained throughout the run. For tertiary, a five year length is used for all countries (during the run).

6.1.1 Data Sources for Student Flows

UNESCO has the most extensive data on all these flow rates. However, as we move up the educational ladder towards secondary and (more so) in tertiary, the data gets skimpy both in terms of cross-sectional, longitudinal and conceptual coverage. Sometimes, World Bank World Development Indicators have a better longitudinal coverage than the published UNESCO sources. For tertiary and secondary, we used data from OECD (generally with a richer country coverage) and NCES (US national Center for Educational Statistics).

6.1.2 Procedure for Initialization with Empirical Data or Estimation

Data gathered from sources mentioned above are stored into Microsoft Access tables inside the IFs historical data file, IfsHistSeries.mdb. These data tables are used to initialize the flow rates according to the following procedures followed in the order they are described here.

i. **Empirical data**: Data values are used to fill-up the values of different flow rates for individual countries, for the year 2000 for the current run and for the year 1960 for historic run of IFs. Most recent (within 10 years of 2000) and earliest (within 20 years of 1960) data values are used if data for 2000 or 1960 were not available. When both of these endeavors to get a data value failed, estimation begins. Two major estimation techniques are used. For some of the education data (e.g. transition rate from primary to secondary), which are conceptually new, the time-series contain only recent data. In such cases, the historic load (1960 as the initial year) is directly initialized from cross-sectional function (see iii below). For at least one variable, tertiary intake rate, two different intake types (described as tertiary type A and tertiary type B by OECD) are combined to get one intake rate. The combination is done by adding the two rates with a possibility of 10% overlap.

ii. **Estimation by trend extrapolation**: In this method, used for some of the variables, initial year value for any country is obtained from a linear regression using available empirical data points for the country. For the 2000 value points after 1985 are only used in regression, while for the 1960 value, all available points are used. For 2000 run, a cross-sectional function might be used in the cases of less than 2 points being available for regression. Also, for 2000 run, in the case of only one data point being available, that data point is priviledged over the regressed value by adding a shift. The routine used for this estimation is titled LReg and coded inside function.bas.

iii. **Estimation from cross-sectional function**: Cross-sectional functions based, mostly, on GDP per capita (PPP) are used to fill holes when no value exists in the data set and the linear extrapolation, when used, returns null. Two separate cross-sectional functions are developed in each case, one for the current (2000) load and one for the historic (1960) load, except in special cases where, in the absence of any historic data (for 1960 or a nearby year), the same function as current load is used for historical load. The cross-sectional functions are, mostly, logarithmic (figure 6.1 below), indicating a faster increase in educational access, participation and achievement as countries get richer and a settling down thereafter towards the natural bounds on those rates. The historic and current functions, in general, have the same shape but the current one is shifted to a better position (fig 6.2). For enrollment rates, the net and gross rates work better than GDP per capita (PPP) as the driver for each other, with appropriate bounds (e.g., 100 percent for net enrollment) [the downward adjustment of gross enrollment, once it has reached its peak and the survival starts to get better, as in the case of primary, cannot be captured by the same cross-sectional function representing the increase in gross enrollment with net].

iv. **Estimation from structural assumptions**: The lack of adequate data for secondary and tertiary is overcome by making assumptions about the secondary and tertiary student flows. For secondary, we have transition rates from primary to secondary, which is applied on primary completion (graduation) rate to obtain a secondary intake rate. We also have sufficient data on secondary enrollment rates (both net and gross). However, we do not have adequate graduation or survival data for secondary (we have some graduation data from OECD). We know that the secondary enrollment will lie somewhere in between the secondary intake and graduation. With an equivalent dropout at each grade, the enrollment will lie just in the middle in between intake and graduation. We thus, assume an equal dropout rate at each grade of secondary and thus, obtain the secondary graduation rate from the enrollment and intake. This might distort the number of students at each grade, but will have no effect on the total enrollment in secondary and the graduation rate at this level, which are the variables we are ultimately concerned with. Similar assumptions are made at the tertiary level not at the pre-processor, but during the model run.

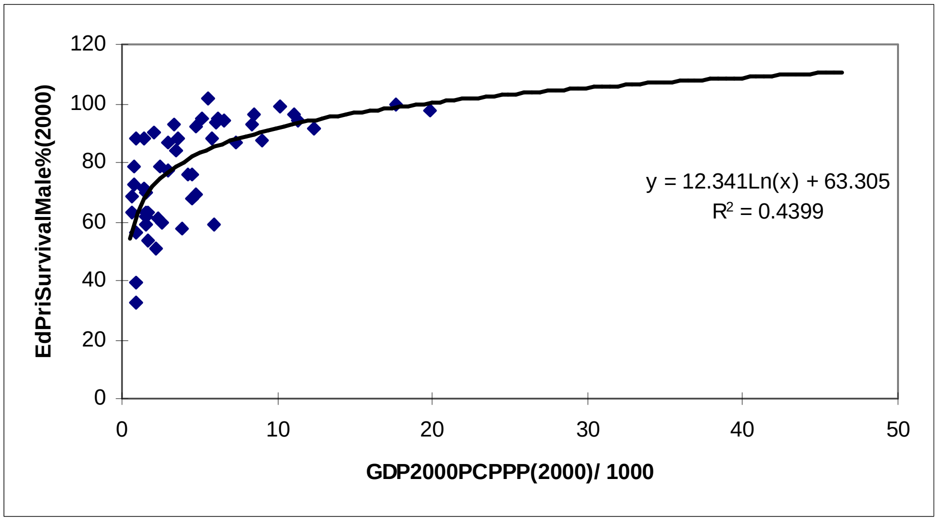


Fig 6.1. Cross-sectional function of survival rate to the last grade of primary (boys) with GDP per capita at PPP

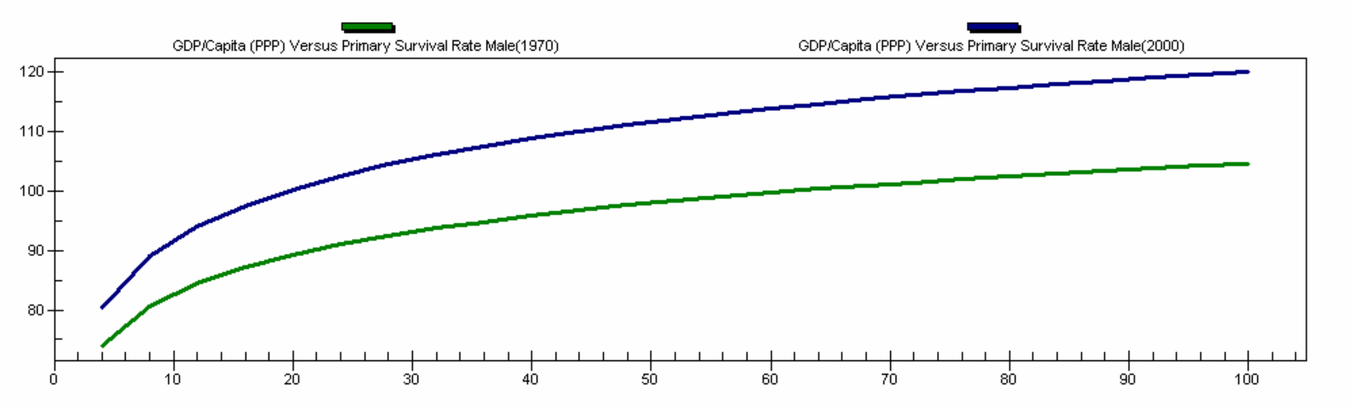


Fig 6.2. Cross-sectional function of survival rate to the last grade of primary (boys) with GDP per capita at PPP, current (2000) and historical (1970)

Here we shall name some of the flow variables, the data tables from which they are read and the cross-sectional functions that are used:

> Primary Gross Intake rate, male ; SeriesEdPriAIRMale%; GDP/Capita (PPP) Versus Primary Apparent (Gross) Intake Rate Male(2000); GDP/Capita (PPP) Versus Primary Apparent (Gross) Intake Rate Male(1970)
>
> Primary Gross Intake rate, female; SeriesEdPriAIRFemale%; GDP/Capita (PPP) Versus Primary Apparent (Gross) Intake Rate Female(2000); GDP/Capita (PPP) Versus Primary Apparent (Gross) Intake Rate Female(1970)
>
> Primary Gross Intake rate, total; SeriesEdPriAIRTotal%; GDP/Capita (PPP) Versus Primary Apparent (Gross) Intake Rate (2000); GDP/Capita (PPP) Versus Primary Apparent (Gross) Intake Rate (1970)
>
> Secondary Net Enrollment Rate, male; SeriesEdSecEnrollNetMale; Secondary Gross Enrollment Versus Secondary Net Enrollment Male (2000); Secondary Gross Enrollment Versus Secondary Net Enrollment Male (1970)

6.1.3 Consistency check on flow rates

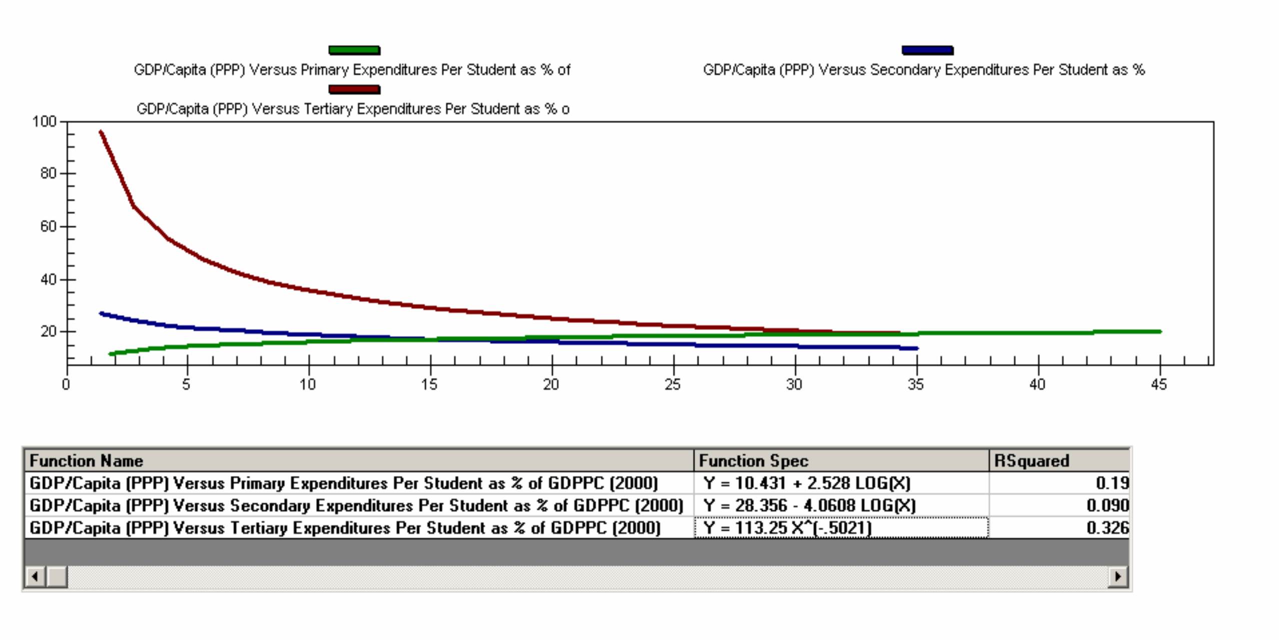
Data inconsistency might arise from the mismatch between an estimated and an empirical value or poor quality data values themselves. Several types of consistency checks are applied to the data initialization described above, so that the values of the flow variables conform with their conceptual definitions. For example, all the net rates are checked against the corresponding gross rates to make sure the net rate is below the gross rate, net rates are further bound at 100 percent, their definitional limit. Similarly, consistency among successive stages of student flow are also maintained (to be specific, intake>enrollment> survival>graduation). Finally, a reconciliation algorithm is used to take care of any gross discrepancy between different flow variables. To elaborate, let us suppose that the historical database recorded (or estimation procedures returned) a primary intake rate and a primary survival rate for a country, which should result in a much higher enrolment rate than that obtained (or estimated) initially by pre-processor. At this point, the reconciliation algorithm, takes all these rates and modifies the least reliable one so that the rates conform.

6.2 Budgetary Flows: Public (Current) Expenditure per Student

Data required for determining the budgetary flows, e.g., public expenditure per student at each level of education, are also initialized from UNESCO time series. The methodology for imputing the missing data is more or less the same as those for imputing the students flow data, the only difference being the non-use of the trend extrapolation.. Expenditure data, quite naturally, are not differentiated by gender at the per pupil level. Another thing that is different in this case is the conversion of per student expenditure from a relative measure (expenditure per student as a percentage of GDP per capita) to a nominal value (in 1995 dollars), so that they can later be used for calculating the total dollar expenditure at each level. The cross-sectional functions are drawn from the relative measures though, since the changes in per students expenditure, the bulk of which goes towards teacher salary, should adequately be reflected by the changes in income level.

Here we shall name the expenditure variables, the data tables from which they are read and the cross-sectional functions that are used:

> Educational expenditures per primary student; SeriesEdExpPri%GDPPC; GDP/Capita (PPP) Versus Primary Expenditures Per Student as % of GDPPC (2000); GDP/Capita (PPP) Versus Primary Expenditures Per Student as % of GDPPC (1970) – Logarithmic Function
>
> Educational expenditures per secondary student; SeriesEdExpSec%GDPPC; GDP/Capita (PPP) Versus Secondary Expenditures Per Student as % of GDPPC (2000); GDP/Capita (PPP) Versus Secondary Expenditures Per Student as % of GDPPC (1970) – Logarithmic Function
>
> Educational expenditures per tertiary student; SeriesEdExpTer%GDPPC; GDP/Capita (PPP) Versus Tertiary Expenditures Per Student as % of GDPPC (2000); GDP/Capita (PPP) Versus Tertiary Expenditures Per Student as % of GDPPC (1970) – Power Function


🟩 GDP/Capita (PPP) Versus Primary Expenditures Per Student as % of
🟦 GDP/Capita (PPP) Versus Secondary Expenditures Per Student as %
🟥 GDP/Capita (PPP) Versus Tertiary Expenditures Per Student as % o
100
80
60
40
20
0
5
10
15
20
25
30
35
40
45

Function Name	Function Spec	RSquared
GDP/Capita (PPP) Versus Primary Expenditures Per Student as % of GDPPC (2000)	Y = 10.431 + 2.528 LOG(X)	0.19
GDP/Capita (PPP) Versus Secondary Expenditures Per Student as % of GDPPC (2000)	Y = 28.356 - 4.0608 LOG(X)	0.090
GDP/Capita (PPP) Versus Tertiary Expenditures Per Student as % of GDPPC (2000)	Y = 113.25 X^(-.5021)	0.326

Fig 6.3: Expenditure (per student) functions

Consistency check of Budgetary Flows

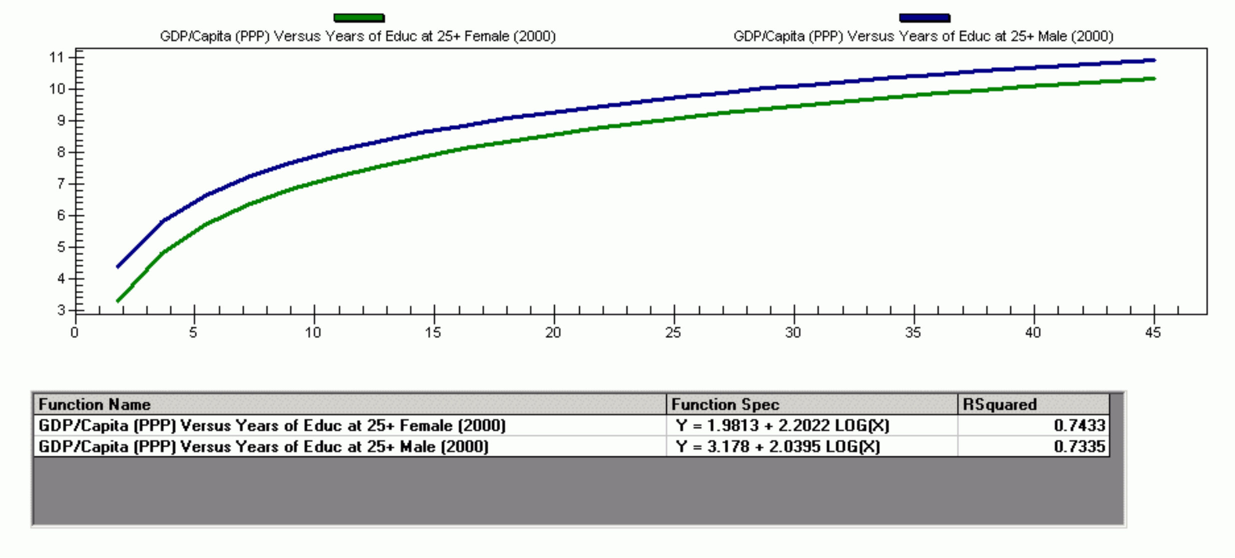
The only consistency checks applied to per student expenditure data are upper and lower bounds imposed on the percentages of GDP per capita spent per student in primary, secondary and tertiary. The upper and lower bounds are obtained by examining the data. While the lower bound is the same (7% of GDP per capita) for all three levels, the upper bounds climb up from 40% for primary to 85% for secondary and to 1000% for tertiary. The high bound is tertiary results from some very low income African countries.

6.3 Human Capital Stock

For human capital stock, the best data set is the Barro-Lee dataset available from University of Harvard Center for International Development. This data set provides the estimates of levels of educational attainment for two different age groups- over age 15 and over age 25 disaggregated by gender at five-year intervals for the years 1960-2000 for a broad number of countries. After some recalculation, e.g., calculating male percentages from female and total data and calculating the percentages of people with at least a specific level of education by subtracting the percentages with a portion of that education, we have collated Barro-Lee data into nine Access tables (three levels, primary, secondary, tertiary x three gender sub dimensions, male, female and total) inside IFs historical database. The education pre-processor has used these data to initialize three different variables, EDPRIPER, EDSECPER and EDTERPER, representing the percentages of adult male or female (15 years and over) population who has completed at least primary, up to secondary or all through tertiary education respectively. For filling up the holes, in this case, three of the procedures described above (i.e., empirical data, trend extrapolation and cross-sectional function) are used in the order listed. We initialize a second human capital stock in the pre-processor, namely the average years of education of adult male and female 25 and over. The data for this one also comes from the Barro-Lee dataset and we use empirical data when available, filling the holes with two cross-sectional functions, one for male and one for female.

Here we shall name the expenditure variables, the data tables from which they are read and the cross-sectional functions that are used:

> Education, primary, percentage of adult (15+) population, male; SeriesEdPriAdultGrads15Male%; GDP/Capita (PPP) Versus Primary Graduates Male (2000); GDP/Capita (PPP) Versus Primary Graduates Male (1960)
>
> Education, primary, percentage of adult (15+) population, female; SeriesEdPriAdultGrads15Female%; GDP/Capita (PPP) Versus Primary Graduates Female (2000); GDP/Capita (PPP) Versus Primary Graduates Female (1960)
>
> Education, years obtained by population 25+, female; SeriesEdYearsAge25Female; GDP/Capita (PPP) Versus Years of Educ at 25+ Female (2000); GDP/Capita (PPP) Versus Years of Educ at 25+ Female (1960)


🟩 GDP/Capita (PPP) Versus Years of Educ at 25+ Female (2000)
🟦 GDP/Capita (PPP) Versus Years of Educ at 25+ Male (2000)
11
10
9
8
7
6
5
4
3
0
5
10
15
20
25
30
35
40
45

Function Name	Function Spec	RSquared
GDP/Capita (PPP) Versus Years of Educ at 25+ Female (2000)	Y = 1.9813 + 2.2022 LOG(X)	0.7433
GDP/Capita (PPP) Versus Years of Educ at 25+ Male (2000)	Y = 3.178 + 2.0395 LOG(X)	0.7335

Fig 6.4: Cross-sectional functions for average years of education among the adult population 25 and over

Consistency check of Human Capital Stock

Lower bounds are imposed on initial values of human capital stock by examining the available data. For some low income countries, the percentages of educated people vary substantially between 1960 and 2000. This is why, we have further differentiated the lower bounds into their historic and current value. The education preprocessor also ensures that the initial year percentages of educated people with a certain level of education stay below the initial year graduation rate at that level. Average years of education is bound between 0.1 to 30 years [should be bound at 20?, mti].

7. Socio-Political

The socio-political module of IFs is relatively simple compared to many of the others. It contains, for instance, no equilibrium-seeking dynamics with supply and demand sides such as those in the agriculture, energy and economic modules. Nor does it contain a substantial stock-and-flow accounting system such as those in population, agriculture, energy, economic, education, and values modules. For the most part, therefore, the socio-political preprocessor needs only to read and process, including the filling of holes, a substantial variety of socio-political variables.

7.1 Global and Regional Variables

The first read in the preprocessor is of a global (non-country specific) parameter determining the direction of the global democracy wave (DEMOCWDIR). That parameter has simply a -1,1 value, specifying whether the wave is ebbing or flowing.

Another variable related to democratization is the swing state. The concept is that certain large emerging countries may influence others in their "neighborhood" with respect to the level of democratization. For an IFs project on the concept, the states so identified were: Brazil, Indonesia, Mexico, Nigeria, Pakistan, Russia, South Africa, Turkey, and the Ukraine. In each case, members of their neighborhoods were identified and stored in a SwingStates table within the IFs.mdb data file. When the values for SwingSts are read, nulls are set at 0, identifying them as not being part of the neighborhood. [Note: move swing state code up above the code below]

7.2 Country-Specific Variables

The preprocessor next moves to a series of country-specific values from data tables:

Freedom from the Freedom House (SeriesFreedom, GDP/Capita (PPP)
Versus Freedom –Logged and Inverted (2005))
Economic freedom (SeriesFreedomEcon, GDP/Capita (PPP) Versus
Economic Freedom (2000) Log)
Governance effectiveness (SeriesGovernanceEffect, GDP/Capita (PPP 2000)
Versus Governance Effectiveness (2002) Linear)
Corruption perception (SeriesCorruption, GDP/Capita (PPP 2000) Versus
TI Corruption Perception (2000) Linear)
Polity project's democracy (SeriesDemoc, GDP/Capita (PPP) Versus
Democracy, Polity measure (1999))
Polity project's autocracy (SeriesAutoc, GDP/Capita (PPP) Versus
Autocracy, Polity measure (1999))
Gender empowerment measure, UNDP (SeriesGEM, GDP/Capita
Versus Global Empowerment Measure) [note name is wrong]
Telephone lines (SeriesTelephoneLines, GDP/Capita (PPP) Versus
Telephone Lines per 1000 (2000))
Size of largest ethnic group (SeriesEthnic1, 0)

Size of second largest ethnic group (SeriesEthnic2, 0)
Percent of population using internet (SeriesInternet%Pop, GDP/Capita (PPP) Versus Internet Percent Use (2000))
Military spending as percent of GDP (SeriesGovtMil%GDP, GDP/Capita Versus Govt Exp Mil as % of GDP (2000) - linear)
Health spending as percent of GDP (SeriesGovtHl%GDP, GDP/Capita (PPP) Versus Govt Exp Hlth as % of GDP (2000) - log)
Public education spending as percent of GDP (SeriesGovtEdPub%GDP, GDP/Capita (PPP) Versus Govt Exp Educ as % of GDP (2002) - log)
Government R&D spending as percent of GNP, OECD (SeriesR&DGovt%GNP, GDP/Capita (PPP) Versus Govt Exp R&D as % of GNP (2000) - linear)
R&D spending as percent of GDP, WDI (SeriesR&D%GDPWDI, GDP/Capita (PPP) Versus R&D Total as% of GNI (2000) - linear)

The measurement of freedom by Freedom House is on a scale in the data that runs from 2 to 14, with lower values being most democratic. This proved consistently confusing for large numbers of users, who expected higher values to be more free/democratic. Therefore the scale in IFs has been inverted by subtracting original values from 16. The function to fill holes has similarly been inverted from the original scale. Values on the inverted variable therefore run 2 to14 (less to more democratic).

Values for governance effectiveness, computed from the function to fill data holes are bound between 0 and 5. Other values from functions are similarly bound. Those for corruption perception are bound between 1 and 10. Those for telephone density are bound between 0 and 1500. Those for Polity’s democracy and autocracy measures are bound between 0 and 10. A combined Polity democracy and autocracy measure (DemocPolity) is computed as democracy minus autocracy plus 10, providing a scale from 0 to 20. The same measure is often used in projects based on the Polity data. The gender empowerment measure is bound between 0 and 1.5.

Moving to the use of the internet, the model needs both an initial condition for the percentage of population using the net in 2000 and the growth rate of that population. To obtain the growth rate, percentage usage rates are read for both 2000 and 2002. In the case of null values for 2002, the percentage is read for 2001, and in case that is null also, a cross-sectionally estimated function, GDP/Capita (PPP) Versus Internet Percent Use (2002), is used to estimate values for 2002. Such a function is also used, as necessary, to fill missing values for 2000. An annualized growth rate is computed from the values for 2000 and 2002. [Probably, if one is computed from the function both should be.]

Government consumption expenditures, in the categories of military, health, education, and R&D spending are dealt with as a block, because the model needs those spending levels converted from percentages of GDP, which is primarily what comes from data, to percentages of government spending. The first step in processing them is to read back into memory the total level of government consumption spending (GCON), which was saved in EcoOutput table of IFs.mdb, during the preprocessing of economic data. Government spending as a percentage of GDP (GExpTot) is based on that total spending.

The processing of particular spending categories will, in general, initially treat them also as a percentage of GDP, so that they can be divided by GExpTot to determine their percentage share of government spending. For instance, military spending is read and nulls are filled on the basis of a cross-sectionally estimated function. Values below 0.01 percent are brought up to 0.01 percent (even Costa Rica, with no official military, has some minimal level of security/military spending).   Then the value of military spending as a share of GDP is divided by the total government spending as a portion of GDP and the result is multiplied by 100 to obtain the percentage share of the military in total government spending. [Total, not governmental spending is probably what is needed in education, health and R&D]

The same process is followed for both health and education spending, both of which are brought up to at least 1 percent of GDP if the data or functional calculation indicate lower levels. One significant complication of data in these areas is that they are for total government spending, not central government spending. Because the government data prepared in the economic preprocessor (both revenue and expenditure data) are for central government [this should be checked again], the health and education spending should be also for the central government only. A very arbitrary division of spending by two is done to reduce values from total to central government.

The most complicated of the expenditures to process is that for R&D, because there are skimpy data and two quite different data sources, the OECD and the World Bank's World Development Indicators (WDI). OECD numbers are assumed to be superior for OECD countries and have the additional virtue of being provided in terms of government R&D instead of total country R&D. Thus they are privileged with respect to all countries in the OECD data set. When values in the OECD data are null, and that is true for almost all non-OECD countries, the flow turns to the WDI values.

WDI values are read when available and holes are filled with a function for all countries without data. The WDI numbers or values filled by function are not, however, used directly to complete the R&D governmental spending values. Instead, a ratio is computed for each of the countries not in the OECD database. The ratio divides expected governmental R&D spending (from the function estimated with OECD data) by total societal R&D spending (from the function estimated with the World Bank data). That ratio multiplies the WDI number or computed value in order to convert the data or estimate of total government spending to an estimate of governmental spending. As with other governmental spending categories, the last step is to divide that calculation by total government spending and multiply by 100 to obtain a percentage value.

Although these four categories (military, health, education, and R&D) make up a good share of direct governmental spending, they are by no means all of it. There has to be room for other spending (which would include infrastructure as well as administrative overhead). The four are summed and if they are more than 99 percent of total government spending, they are reduced proportionately. Other spending is computed as a residual in any case.

The final step with respect to governmental spending is to convert all five spending categories (including other) back to absolute spending values by multiplying the percentages times the total governmental consumption.

The flow moves next to conventional and nuclear power initializations. An assumption is made that there is less developed countries obtain more conventional power for the same spending level in dollar terms, a phenomenon equivalent to the distinction between purchasing power parity and market exchange rates. In fact, the global parameter that indicates the greater power return on expenditures for poorer countries (CPOWLDCF) normally has a value for the poorest of 5, about the ratio of PPP to MER GDP per capita values for the poorest countries. That factor is assumed to erode as per capita GDP rises to $10,000 and the scaled value of it is placed into a conventional power factor (CPOWF) for use in the forecasting side of the model.

Moving to nuclear power, the model reads an estimate of nuclear power, expressed roughly in terms of strategic warheads, from data, filling missing values with zeros.

Nuclear power (SeriesNPOW, 0)

Change in nuclear power over time will depend on the share of military spending devoted to it (NMILF). That share is computed by doing a very rough calculation of the annual spending on warheads (their number times .002, which can interpreted as $20 million/warhead with an average life of 10 years; in reality, the nuclear side also involves tactical warheads, delivery systems and more, so this is a very crude proxy of total nuclear spending). The estimate of annual nuclear-related spending divided by total military spending provides a share of the total being devoted to the nuclear side of the military budget. It is bound at no more than 15 percent.

The final initial condition prepared for the model is a specification of European Union membership, with values of 1 indicating membership and 0 indicating non-membership. The values are read from a table within the IFs.mdb file.

The rest of the preprocessing for socio-political variables involves writing the above values to appropriate output files for aggregation into regions as appropriate and for use in the initialization of the model's run file.

8. International Conflict

The preprocessor for international conflict data handles relatively few variables for the model. The somewhat unique aspect of it, however, is that the variables are generally dyadic, that is country-to-country, rather than single country in character. This routine of the preprocessor calls three separate routines for three different dyadic variables: contiguity, territorial disputes, and threat levels.

8.1 Contiguity

The first variable processed is a measurement of contiguity (closeness) of countries (CONTIGUITY). The table for contiguity data is in the IFs.MDB file. Values on the contiguity scale run from 1 (neighboring) to 6 (distant). Most pairs of countries are distant and the file does not contain any specific representation of them. Thus the matrix of all countries with all others is set initially to be full of values of 6, using Paul Diehl's scale. As the data file is processed, other values are found for the dyads in it and the value of 6 for particular cells is replaced as appropriate. The diagonal (countries with themselves) is filled with zeros.

The bulk of the processing is associated with the computation of values for regions. Members of each region are all processed relative to members of each other region. The lowest contiguity value characterizing any dyad of countries across regions becomes the region-to-region value.

8.2 Territorial Disputes

The second routine called and variable processed is territorial disputes. This variable can only have one of two values, 0 for no dispute and 1 for dispute. The matrix is filled with 0s; any dyad found in the table (within the IFs.mdb file) will be a one and will be placed into the dyad. Again, the processing of regions is more complicated and time consuming. If any dyad of countries across two regions has a dispute, the pair of regions is identifying as having a dispute.

8.3 Threat

The third routine handles the computation of threat between countries and then between regions to which those countries belong. The threat data for initialization of values in each dyad are in a file of their own called MIDsHistoryProbabilities.csv under the MIDsInitialValues subdirectory of the Data directory. The data come from the militarized international disputes (MIDs) dataset of the Correlates of War (COW) project and were substantially enhanced in work done by Mark Crescenzi before their inclusion in IFs. There are values for all countries with all other countries; for instance, the dataset includes a 0 value for Honduras relative to Jordan, unlike other data sets in the part of the preprocessor, which would simply have excluded that dyad. Therefore all values are read and put into the appropriate dyad without any need to provide a default value for the dyad. Still again, the primary task is the computation of region-to-region values. Those are calculated as average values across the members of the regions. That is, dispute values are summed for all pairings of countries between two regions and the sum is divided by the numbers of such pairings.

9. State Failure

State failure data come from the work of Ted Robert Gurr and associates via the dataset of the State Failure Task Force (now the Political Instability Task Force). That dataset identifies four types of state failure: ethnic war, genocide/politicide, adverse regime transition, and revolutionary war. In addition, there is a representation of consolidated events, that is any one or more of the four types of failure in a given year. The first year of events in the dataset is 1955.

The data set as prepared for processing by IFs includes tables of event occurrence (0=no event, 1=event) and event magnitude (short scales) across all four event types and for consolidated events. These are, in essence, the raw data on state failure that are available to the preprocessor. The preprocessor works with those raw data to create several collapsed and aggregated tables of data.

First, it was recommended to the IFs project by Ted Gurr that the four types be collapsed into two: internal war (ethnic war, genocide/politicide, and revolutionary war) and domestic instability (adverse regime transition).

Second, three sets of additional tables were prepared for those two collapsed state failure categories and for consolidated events. Those sets represent (1) event probability (capturing the average numbers of events per year, new or continuing, over the most recent 10, 20, 30, and full time period horizons), (2) year one or new event probability (over the same alternative period horizons), and (3) average magnitude (the average annual magnitude of events over the same alternative period horizons). That is, the basic event and magnitude data are processed and put back into the historical data file (IFsHistSeries.mdb) for use in cross-sectional or longitudinal analysis.

The process beings by opening tables for both the raw data and the processed values:

Ethnic war magnitude (SeriesSFEthnicWarMag, 0.0)
Genocide/politicide magnitude (SeriesSFGenocideMag, 0.0)
Adverse regime change magnitude (SeriesSFRegTranMag, 0.0)
Revolutionary war magnitude (SeriesSFRevolWarMag, 0.0)
Consolidated event magnitude (SerieSFsConsolidatedMag, 0.0)

Ethnic war event (SeriesSFEthnicWarEv, 0.0)
Genocide/politicide event (SeriesSFGenocideEv, 0.0)
Adverse regime change event (SeriesSFRegTranEv, 0.0)
Revolutionary war event (SeriesSFRevolWarEv, 0.0)
Consolidated event (SeriesSFConsolidatedEv, 0.0)

Internal war event probability (SeriesSFInternalWarEvProb, 0.0)
Domestic instability event probability (SeriesSFDomInstabilityEvProb, 0.0)
Consolidated event probability (SeriesSFConsolidatedEvProb, 0.0)

Internal war year 1 event probability (SeriesSFInternalWarY1Prob, 0.0)
Domestic instability year 1 event probability (SeriesSFDomInstabilityY1Prob, 0.0)
Consolidated year 1 event probability (SeriesSFConsolidatedY1Prob, 0.0)

Internal war average magnitude (SeriesSFInternalWarMagAv, 0.0)
Domestic instability average magnitude (SeriesSFDomInstability MagAv, 0.0)
Consolidated average magnitude (SeriesSFConsolidated MagAv, 0.0)

The first step in actually processing computes the 10, 20, 30, and all year event magnitude values for internal war (the sum of three event types), domestic instability, and consolidated events. The second step computes the probability of initial or continuing events for the same three categories across the same four different time horizons. And the third step computes the probability of initial events only for the same three categories across the same four different time horizons. Although the blocks of code for each of these steps are fairly long, the process is simply counting, summing, and averaging.

After these three blocks of three variables for various time horizons are complete, the values are written to the nine tables already opened (see the last nine in the list of tables shown above).

The final step in the preprocessor is the writing of 10 variables to the output table of IFs.mdb for actual use in the model. Five are for domestic instability and five are for internal war. The five for each type of collapsed event are: event onset or continuation, event continuation only, magnitude, event probability (new or continuing) over the long or full history and event probability (new or continuing) over the most recent 10 years. The last two "variables," because they are representations of historic events, do not actually vary over time in the model and are used for analysis purposes only. The first three are the key variables that the model forecasts (across the two collapsed event types).

10. Values

The values representation in IFs is closely tied to the two value dimensions that the World Values Survey (WVS) project has identified, analyzed, and rooted empirically in several waves of global surveys. Those two dimensions are (1) traditional versus secular-rational values (TRADSRAT) and (2) survival versus self-expression values (SURVSE). In addition, the WVS created a materialist versus post-materialist index (MATPOSTR) that cuts across those two dimensions. In each case, within IFs there are additive parameters that allow the user of the model (tradsratadd, survseadd, and matpostradd) to posit changes in the values of those indices beyond the values computed in the model; but the values in the base case are all zero and the only duty of the preprocessor with respect to them is to set those zero values.

The WVS has found that global cultural regions do exist. Therefore it is important for the model to have initial values not only on the two orthogonal dimensions and the materialist/post-materialist index, but on the 11 cultural regions (10 named regions and a residual or other category). For instance, it is essential to know the cultural region membership of each country (CULTREG) and to have initial conditions for the shift in value positions on the two dimensions and the cross-cutting index associated with each cultural region (cultshts, cultshse, cultshmp).

Professor Ron Inglehart, the founder of the WVS, generously provided extensive data from the first four waves of the WVS for the database in IFs. A separate file, IFsWVSCohort contains tables on the key dimensions and index as well as on many individual questions. Most of the time, the data are available by age cohort using the 6-cohort division of the WVS.

The preprocessor routine for initialization of the above variables has been used also for the extended analysis of the accuracy of the various formulations for predicting positions of countries on the value dimensions, computing the values anticipated by the formulations, comparing them with actual values, and putting the differences (called loads) into parameters that are available for display in the model but that are not changed in it (or used by it). For instance, the parameter TRADSRATLD for the United States carries a substantial negative value, because the US is more traditional than would be expected by the formulations used to forecast positions on that dimension. Similarly, the preprocessor computes such differences (loads) by cultural region to indicate the manner in which entire groupings of countries differ from what would be expected in the formulations. For instance, the parameters CULTSHTS and CULTSHTSLOAD (the first using a multivariate function and the other a simple mean) for the English speaking cultural region also carry substantial negative (and identical) values, because that region as a whole is more traditional than would be expected based on variables such as GDP per capita or labor force in various sectors.

Further, the preprocessor has been used to attempt forecasting of country positions on specific questions (such as trust in others or happiness). These incremental features have nothing to do with preparation of initial conditions for the model and will be discussed only in passing – unlike the load parameters above, the results are put by the preprocessor into files that are not even seen by the model users.

The preprocessor begins with looking to data tables for data of three different kinds. The first, taken from IFsWVSCohort, are data on the two dimensions and the summary index, necessary for initializing the model: [why haven't we gone to W4, which is available in the data for cohort data, but not loads?]

Survival/Self-Expression, by cohort (SurvSelfExpW3, GDP/Capita (PPP) Versus Survival/Self-Expression Cohort x)
Survival/Self-Expression loads, no cohort, (SurvSelfExpLoadsLastestW3, GDP/Capita (PPP) Versus Survival/Self-Expression Total)
Traditional/Secular-Rational, by cohort (TradSecRatW3, GDP/Capita (PPP) Versus Traditional/Secular-Rational Cohort x)
Traditional/Secular-Rational loads, no cohort, (TradSecRatLoadsLastestW3, GDP/Capita (PPP) Versus Traditional/Secular Rational Total)

The second type of data contains values on individual questions, also taken from IFsWVSCohort. As indicated, these are for project analysis and not for initialization of the model:

Percent happy (HappyPercentW3)
Trust in people (TrustPeoplePercentW3)
Respect of authority (RespectAuthorityPercentW3)
Have signed petition (SignedPetitPercentW3)
God is important or them(GodImprtPercentW3)
Homosexuality is (never) justified (HomoJustifPercentLatest)
Abortion is (never) justified (AbortJustifPercentW3)
Proud of their nation (NationProudPercentW3)
? (PMMinMatPercentW3)
Autonomous personally (AutonPercentW3)

A third type of data holds key driver variables for analysis of change in position on the dimensions and for the computation of the country-specific and cultural region load parameters discussed above. These come from the IFsHistSeries.mdb database:

Portion of labor in industry (SeriesLaborInd%, 25)
Portion of labor in services (SeriesLaborSer%, 50)
Whether country is ex-communist (SEriesCulExComDum, 0)

In addition, one variable, the position on the materialist/post-materialist dimension, is taken from the WVSCohortWave2 table of IFs.mdb: [why is this only wave 2?]

Materialist/postmaterialist (Social, GDP/Capita (PPP) Versus Materialism/ Postmaterialism Cohort x [1-6])

Finally, one parameter, the cultural region of each country is taken from the Social table of IFs.mdb:

Cultural region of countries (CultReg, Other)

The first actual reading and preparing of data is a load of the cultural region location for all countries (CULTREG). A count is also maintained of the number of countries for which data are available.

The next read is of values on the materialism/post-materialism index for all six of the cohorts and the country total. Holes are filled with cohort-specific functions estimated cross-sectionally. Values are bound between 1 and 5.

Then values for a cultural shift calculation on materialism/post-materialism are computed as the difference between the values read and those predicted by the cross-sectionally estimated function. [This block looks very iffy – the variable cultshmp is used on the right side before it has been given any value; the calculation is only for the first cohort; later on the variable is apparently calculated again, suggesting that this is dead code]

The next read is of values on the survival/self-expression dimension for all six cohorts and the country total. Holes are filled from a function and values are bound between -2 and 2.

The next three blocks of code read and process variables relevant to the calculation of shifts to the values of countries on the survival/self-expression dimension related to cultural region membership. Specifically, the variables used in predicting factor loadings are read (the percentage of labor in industry and services and the ex-communist state dummy variable). A formula from work with Ronald Inglehart is used to predict factor loadings on the survival/self-expression dimension. The deviations of predictions from values in the dataset are summed up by cultural region (across all countries in the cultural regions for which there is data). [two predictions are made, by function and by mean; they produce ultimately the same shift and load parameters for all cultural regions – why? Seems to be simply an internal check]

The next read is of values on the tradition/secular-rational dimension for all six cohorts and the country total. Holes are filled from a function and values are bound between -2 and 2.

The next two blocks of code are for computations of culture-region shift factors. First, functions from work with Ronald Inglehart are used to compute predicted values for all countries. Then, for all countries with data as well, the differences between predicted and actual values are summed within their respective cultural regions.

At this point the absolute magnitudes of net shifts on both dimensions and the index have been summed for all cultural regions, using only members of cultural regions with data. Thus division by the number of countries within cultural regions with data can provide the average parametric shift that is associated with cultural region membership. These are computed based on the more complex equations (CULTSHMP, CULTSHSE, CULTSHTS) and based on means only (CULTSHMPLOAD, CULTSHSELOAD, CULTSHTSLOAD). As indicated, these are not used in the model, but are of interest because they indicate how far individual countries or cultural regions are from the values that would be expected based on predictive functions.

The remainder of the routine computes expected values for countries on a selection of important questions form the World Value Survey. The questions for which data are read and compared with the predictions were indicated earlier. The original purpose of this work was to forecast wave 4 values from wave 3 data, in order to explore the proposition that changes in values over time can be forecast with some degree of accuracy (Inglehart and Welzel subsequently pursued this further). The values for the forecast are written to files and not prepared for use in the model itself.

11. Conclusions

The data preprocessor of IFs has dramatically proven its worth. It has made it possible for the project to regularly and quite easily translate updates of the database, both major and minor ones, into new initial conditions for the model runs. It has facilitated the addition of new countries and the flexibility of regionalization across countries. It has also made possible the simultaneous creation and use of historical and future-oriented data loads.

Although there are many rough edges in the code of the preprocessor, including important and somewhat arbitrary assumptions that could be better backed up with data and research, it provides a basis for constant incremental enhancements and periodic significant additions to the model (such as new modules).

Few users of IFs even know that the preprocessor exists, but the concept of it and the development of it have been of great importance for the IFs project. Long may it live and prosper.

Bibliography

Chenery, Hollis. 1979. *Structural Change and Development Policy*. Baltimore: Johns Hopkins University Press.

Hossain, Anwar with Barry B. Hughes. 2004 (July). The Database of International Futures. Unpublished IFs working paper on the IFs website.

Hughes, Barry B. and Anwar Hossain. 2003 (September). Long-Term Socio-Economic Modeling. Unpublished IFs working paper on the IFs website.

Hughes, Barry B. with Anwar Hossain and Mohammod T. Irfan. 2004 (May). “The Structure of IFs,” Unpublished IFs working paper on the IFs website.

Hughes, Barry B. and Evan E. Hillebrand. 2006. *Exploring and Shaping International Futures.* Boulder, Co: Paradigm Publications.

Systems Analysis Research Unit (SARU). 1977. *SARUM 76 Global Modeling Project*. Departments of the Environment and Transport, 2 Marsham Street, London, 3WIP 3EB.