GTAP 5: A Large-Scale Data Base Construction Project

Betina V. Dimaranan and Robert A. McDougall[1]

June 2000

ABSTRACT

A publicly available, fully-documented, global data base is the centerpiece of the Global Trade Analysis Project. Since GTAP's inception in 1993, the data base, along with the other components of the Project, have continued to evolve and grow in response to and with the support of the users of the data base. The regional classification has expanded from an initial 15 regions in the first version to 65 regions in the first pre-release of GTAP 5. The 57-sector classification is a result of increased sectoral disaggregation especially in the agricultural and food sectors and, more recently, in the service sectors. The quality and processing of data inputs, such as the domestic data bases, the bilateral trade data, the protection data, and the energy data, continue to be improved on from one version of the data base to the next. Over the last seven years, the GTAP data base has supported quantitative economic analysis using the GTAP model and other multi-regional, applied general equilibrium models.

As the regional and sectoral classification expands and as new features are added to the data base, the size and complexity of the data base construction task also increases. New methods and standards have been adapted in order to facilitate complete automation of the process, ease in replicating previous builds, and flexibility in adding new inputs or regions in the data base. These methods and standards are introduced in building the GTAP 5 data base. They include the use of a modular structure, build management using a recursive `make` procedure, version control for the program files and input data files, regional flexibility, and use of a common tool set. Although further improvements have to be done, these innovations in the data base construction process has already allowed for almost completely automation, replicability and regional flexibility in the construction of the GTAP 5 (pre-release) data base.

[1] Post-Doctoral Fellow and Deputy Director, respectively, at the Center for Global Trade Analysis, Purdue University, West Lafayette, Indiana 47907, USA.

GTAP 5: A Large-Scale Data Construction Project

1 INTRODUCTION

The Global Trade Analysis Project (GTAP) is comprised of a publicly available, fully documented, global data base, a standard modeling framework and associated software, a global network of researchers, and a consortium of national and international agencies that provides leadership and base-level support to the project. Since its inception in 1993, the project and its components have continued to evolve and grow. The standard modeling framework continues to be improved on with new features (e.g. welfare decomposition) and extensions (e.g. imperfect competition, dynamic GTAP model) as well as improvements in the associated software (e.g. RunGTAP). The number of researchers attracted to the project and linked via the GTAP mailing list or the GTAP web site continue to increase. Consortium membership, which has risen from 2 agencies in 1994 to18 national and international agencies to date, attests to the increased importance of the project in the area of global economic analysis. The project also continues to offer an annual short course, more recently with a distance-learning, preparatory course component, and for the past three years has sponsored an annual conference on global economic analysis.

The key component of the project, production of a global data base, has also seen significant improvement over the last seven years. The GTAP data base that is scheduled to be released in this fall is already the fifth version since1993. With the disaggregation composed of at least of 65 regions and 57 sectors, along with better quantified data inputs and the improved construction procedure, is now a much bigger and better data base compared to the 24-region by 37-sector version 2 data base released in 1994. The data base is produced through the joint effort of external data contributors, including some individuals from the consortium member agencies, and staff at the Center for Global Trade Analysis (CGTA). Typically, the external data contributors provide domestic data bases (input-output tables) and international data sets such as trade data and protection data while staff at the CGTA handle the data base construction and assembly.

A new version of the GTAP data base is produced approximately every eighteen months. Under this cycle, six months are devoted to developing the requirements for the data base construction and securing data from external contributors. Another six months are allocated for the preparation of a pre-release data base which is made available to the consortium members. A final data base is made publicly available six months after the pre-release. The pre-release data base or, starting with version 5, the pre-releases of the data base are made available only to members of the GTAP consortium. This policy gives the consortium members early access to the data base and also provides time for identification and correction of errors before it is made available for wider distribution. Individuals who contribute domestic data bases receive both an aggregation of the pre-release data and the complete final data base.

The GTAP data base supports quantitative economic analysis not only in conjunction with the GTAP model but also with other multi-region models. This paper aims to provide the reader with an appreciation of the GTAP data base, its historical development, and the complex process involved in bringing together the domestic data bases and several international data sets to create a global data base of value flows describing the world economy for a given benchmark year. It also aims to inform the interested reader about the recent innovations that have been adapted at the Center in order to make the data base construction process fully automated, replicable and flexible.

A brief documentation of the historical development of the GTAP data base, from the first version to the current publicly available version, GTAP 4, is given in the next section. The third section of the paper describes the features of the forthcoming GTAP 5 data base. The data base construction process is outlined in the fourth section of the paper before the recent innovations to the data base development process are discussed in the fifth section. Some concluding remarks are provided in the final section of the paper.

2 HISTORICAL DEVELOPMENT OF THE GTAP DATA BASE

The GTAP data base consists of bilateral trade, transport, and protection matrices that link the economic data bases of individual countries or regions. It is constructed from a collection of input-output tables from individual countries (domestic data bases) and international data sets. The basic procedure is to start with the input-output tables and update them with macroeconomic, bilateral trade, energy, and protection data and assemble the parameters file. The increasing use of the data base in quantitative economic analysis has spurred the demand for greater disaggregation, new features, and better quantification of critical data components. The characteristics of the previous GTAP data bases, particularly the new or revised features of each version, are discussed in this section.

2.1 The First Two Versions

The earliest version of the data base relied on input-output tables from the SALTER project of the Australian Industry Commission. This included 15 source input-output tables. The SALTER project, undertaken in the 1980s and early 1990s, had a greater disaggregation of the Asia-Pacific region reflecting their focus on issues in that region. GTAP also adopted the SALTER sectoral classification which identified 37 commodities. The data construction was done using much of the software and procedures from the SALTER project as well.

The domestic data bases were updated using new bilateral trade and protection data. Most of the protection data work was done by Bradley McDonald who was then with ERS/USDA. Tariff data was drawn from individual country Trade Policy Reviews conducted by the then General Agreement on Tariffs and Trade (GATT). Agricultural support and protection data were obtained from country studies of Producer Subsidy Equivalents (PSEs) undertaken by the OECD and ERS/USDA. This initial data base was made available only to a few individuals and was not publicly released.

In the second version (GTAP 2), the 37-sector classification was maintained but some SALTER domestic data bases were updated and new regions were added to make a total of 24 regions. A list of sectors and a list of regions for versions 2 to 4 are given in tables 1 and 2, respectively. New macroeconomic data was introduced in order to update the benchmark year from 1990 to 1992. New protection data was also incorporated in the Version 2 data base. This includes improved tariff data which were based on tariff schedule submitted by country members to the WTO under the Uruguay Round negotiations. The disaggregated tariff data, provided to GTAP by the US Trade Representative’s office, was aggregated from the tariff line level using import weights. Estimates of non-tariff barriers including anti-dumping duties, countervailing duties, price undertakings, and export restraints on textiles and wearing apparel, were also included in the protection data set.

Policies related to the availability of the GTAP data base to potential users were instituted when GTAP 2 was publicly release in 1994. The full GTAP data base was made available at different price levels which depend on whether the subscriber is a government/private sector user, a multiple academic user, or a single academic user. Aggregations of the data base were made available for a small fee. The aggregations, limited to within a 10-region by 10-sector dimension, were prepared for the subscriber and also made available to the public via the GTAP ftp site (and later the GTAP web site). The version 2 data base is the one used in the simulations performed for the studies that were included in the GTAP book (Hertel, 1995).

2.2 GTAP 3 and Full Documentation

Aside from significant changes introduced in the GTAP 3 data base in the areas of regional disaggregation, trade, and protection, perhaps the more significant of the changes were in the increased professionalization of data base construction and management and also in documenting the data base. In addition to the documentation on the preparation of the different data inputs and the procedure in assembling the global data base, McDougall (1997) also provides some summary tables on many interesting aspects of the database, including GDP shares, cost shares, input-output multipliers, and effective rates of protection.

While the 1992 benchmark year and the 37-sector classification of the version 2 data base were maintained, the regional classification was increased from 24 to 30 in the third version of the data base. India was broken out from the South Asia region; Chile, EU3, and EFTA were introduced as new regions; and the Economies in Transition (EIT) region of GTAP 2 was split into the Central European Associates (CEA) and the Former Soviet Union (FSU). In addition to input-output tables for the new regions, some older input-output tables were also updated.

Work on the estimation of bilateral merchandise trade data was begun by Marinos Tsigas at Purdue in the late 1980s. Continuing on this work, Mark Gehlhar has refined the estimation and reconciliation of divergent bilateral trade flows. For version 3, Gehlhar started reconciling bilateral flows at a disaggregate level before they are aggregated up to GTAP commodity classification (Gehlhar, 1997). For trade in non-factor services, data used in versions 3 and 4 of the data base came from several international institutions and from Alan Fox (based on his work on the Michigan Model with Deardorff and Stern). The external data served as the starting point for estimating the trade flows of non-factor services. Specifically, as RAS procedure is used to match the data to a GTAP country/commodity concordance and to match the target total trade flows from World Bank data. Estimates of the composition of total exports and imports of services, obtained from the individual country input-output tables, are also used as inputs in the procedure.

For protection data, GTAP 3 benefitted from work done at the World Bank for a 1995 conference on the Uruguay Round and the Developing Countries (Martin and Winters, 1996). This includes estimates of pre- and post-Uruguay Round protection as compiled at the World Bank from the GATT Integrated Data Base and from other sources (Reincke, 1997; Ingco, 1997). Aside from the pre- and post-UR import tariff estimates, the version 4 protection data included estimates of export subsidy expenditures and well as information on non-tariff barriers outside of agriculture. These include estimates of antidumping duties for the USA, Canada and EU; price undertakings for the EU, and some export tax equivalents associated with voluntary export restraints (VERs) and with the Multifiber Agreement (MFA). With the richness and timeliness of the protection data incorporated in the GTAP 3 data base released in 1996, it was widely used in analyzing the various impacts of the Uruguay Round.

The parameters file contains the behavioral parameters used by the GTAP model. As in previous versions of the data base, the source substitution elasticities (Armington elasticities) and factor substitution elasticities in GTAP 3 were adapted from the parameters in the SALTER project. For the price and income elasticities that are used to calibrate the expansion and substitution parameters in the constant difference of elasticities (CDE) utility function that characterizes household demand in the GTAP model, revised estimates were obtained from the literature (Dimaranan, McDougall and Hertel, 1997).

2.3 GTAP 4 and Sectoral Disaggregation

In the GTAP version 4 data base publicly released in the fall of 1998, the sectoral classification was expanded for the first time from the 37 SALTER sectors to 50 sectors, with much of the additional detail going into food and agricultural sectors. Also broken out were motor vehicles and parts and electronic equipment, in light of their dominance in world trade. The utilities sectors – electricity, gas manufacture and distribution, and water – were also disaggregated in order to serve the interest of those working on energy-environment issues. Some new input-output tables were contributed with the desired sectoral detail. However, for most of the tables, especially the older input-output tables, the disaggregation had to be done at the Center. The procedure used in disaggregating the input-output tables to the 50-sector classification in GTAP 4 is discussed under a later section on data base construction procedure and documented in Liu and McDougall (1998).

The regional classification was further expanded from 37 to 45 regions. The expansion includes the further disaggregation of the European Union into United Kingdom, Germany, Denmark, Sweden, Finland, and the Rest of the European Union. Sub-Saharan Africa was split into the South African Customs Union (saf), Rest of Southern Africa (rsa) and the Rest of Sub-Saharan Africa (rss). The other new regions introduced in GTAP 4 are Vietnam, Sri Lanka, Venezuela, Colombia, Uruguay, Turkey, and Morocco. The old input-output tables for some regions, e.g. New Zealand, China, Philippines, Taiwan and Canada, were also updated. Full documentation on GTAP 4 is provided in McDougall, Elbehri and Truong (1998).

Macroeconomic data for 1995 was obtained from the World Bank to support the updating of the reference year from1992 in GTAP 3 to 1995 in GTAP 4. The macroeconomic data requirements of data base construction include GDP, population, private consumption, government consumption, gross fixed investment, capital stock, and depreciation.

Tariff information in the version 4 data base was taken from the UNCTAD TRAINS data and supplied to GTAP by Jersey Rozanski and Emiko Fukase of the International Trade Division of the World Bank. For importing countries which were not covered in this protection data set, tariff information from the version 3 data was used. The post-Uruguay Round tariff estimates in version 3 were dropped.

Significant changes were introduced in the agricultural protection data for version 4. For agricultural protection data, Marinos Tsigas of the ERS/USDA contributed 1995 market price support and subsidy information based on the PSE/CSE data calculated by the OECD. Unfortunately, this data is only for OECD countries and Central and Eastern European countries. Hence, agricultural import protection from the version 3 data base, adjusted for changes in world prices, was adapted for the non-OECD countries. This price-comparison approach was also implemented for export subsidies in agriculture instead of the previous approach, followed in versions 2 and 3, of using country submissions to the WTO on export subsidy expenditures in the Uruguay Round. This is to avoid potential inconsistencies with the price comparison-based import protection data and also because adjusting the export subsidies from the 1992 base period was problematic.

With the exception of the MFA, very little information on non-agricultural, non-tariff barriers were obtained for the version 4 data base. Although it wasn't a desirable option since NTBs are recognized to be transitory in nature, the outdated 1992 estimates of anti-dumping duties and price-undertakings from the version 3 data base were carried along in the 1995 version 4 data base. New estimates of export tax equivalents of the quotas on textiles and clothing under the MFA, from data obtained from Linda Linkins of the USITC and Will Martin of the World Bank, were incorporated in the version data base.

The calculation of energy production targets used in updating the input-output tables (in the FIT process described in a later section), was revised in GTAP 4. Energy production values were calculated based on country-level energy volumes and price data bases. The work done on this by Gerard Malcolm (Malcolm,1998) served as a basis for further work in quantifying the energy data for GTAP 5.

Another area where significant changes have been introduced in GTAP 4 is that of primary factor splits, i.e. how value-added for a given sector is split between land, labor, and capital. For agriculture, the split between payment to factors land, labor, and capital in agriculture are obtained from a survey of the primary factor cost shares from the literature (Tsigas and Hertel, 1997). A distinction between skilled labor and unskilled labor used in each sector in introduced in GTAP 4. Estimates of the labor splits for each region and sector in GTAP 4 were generated from a regression model fitted using labor data from 15 national labor surveys as input (Liu, et.al, 1998). A natural resource sector endowment factor, which applies only to the forestry, fisheries, coal, oil, gas, and other minerals sectors, is also introduced for the first time in GTAP 4. The cost share of this factor for each region was determined as the share which when combined with the elasticities of substitution in valued-added in the model, replicates a target elasticity of supply for each of the resource constrained sectors (Hertel and Tsigas, 1998).

Other significant changes in GTAP 4 include the revised CDE calibration procedure, the revised aggregation program, and changes in the data base construction process. A new method for CDE calibration based on a maximum entropy approach (Liu, et.al, 1998) replaced the constrained minimization approach followed in previous versions of the data base (Dimaranan, et. al, 1997). Unlike the previous versions of the DOS command-line aggregation suite, the package provided with the full GTAP 4 data base excluded the CDE calibration procedure from the aggregation process. This change significantly reduced the time it takes to create an aggregation. The modular structure, as well as the partial automation of the construction process, was introduced in GTAP 4. This laid the groundwork for further improvements in the data base construction procedure in GTAP 5.

Other new features associated with the GTAP 4 data base include the time series trade data and a tax adjustment program. For GTAP 4, the time series data for bilateral merchandise trade assembled by Mark Gehlhar for the GTAP regions and merchandise sectors, covering the years 1965-1995, is provided with the GTAP 4 data base package. A new program developed by Gerard Malcolm for use in making adjustments to the tax/subsidy rates in the data base was introduced after the release of GTAP 4. This program, called ALTERTAX, allows the user to adjust the value flows associated with the tax/subsidy data while keeping the other value flows in the data base mostly unchanged (Malcolm, 1998).

Another major innovation which was first released carrying GTAP 4 data is the Windows-based GTAP data aggregation software called GTAPAgg developed by Mark Horridge of Monash University. The GTAPAgg package included an encrypted version of the full data base and allowed users to make any number of aggregations of the data base each limited to a 10-region by 10-sector aggregation. With the use of a special license file, this limit is relaxed for subscribers to the full data base. The time series bilateral trade data is included in the GTAPAgg package.

3 FEATURES OF THE VERSION 5 DATA BASE

The more significant of the new features in the version 5 data base are the revised regional aggregation, the disaggregation of the services sectors and improved services trade data, and improved data on the energy sectors. These aspects of the data base are described below.The benchmark year for GTAP 5 is 1997, updated from 1995 in version 4. The first pre-release of GTAP 5 was just made available to the consortium members at the time of this writing (May 2000).

3.1 Regional Disaggregation

The two major directions in terms of regional disaggregation for version 5 are the full disaggregation of the European Union and the increased disaggregation of the Southern African region. With funding from the European Commission, the LEI undertook the disaggregation of the EU-15. The first pre-release has 14 EU regions with Belgium and Luxembourg together. Seven Southern African countries were added to the data base as a result of a project conducted by Mark Horridge, Channing Arndt, David Evans, with funding from DIFID. Domestic data bases for Bangladesh, Peru, Hungary, and Poland were also obtained from individual contributors. New input-output tables for Japan and South Korea replaced the older SALTER tables.

The first pre-release of GTAP 5 is composed of the 65 regions and 57 sectors listed in Table 3. The regional disaggregation, as well as some data inputs, will still be updated before the final release of the version 5 data base. This includes updated input-output tables for China, Taiwan, India and Colombia and separate IO tables for Belgium and Luxembourg. Egypt and Tunisia will likely be introduced as new regions in the final GTAP 5 data base.

3.2 Services Sectors and Services Trade Data

As part of a major initiative to provide a better representation of the services sectors and of trade in non-factor services in the data base, improved services trade data was obtained, the procedure for processing the data was revised in a major way, and the services sectors were disaggregated from 8 sectors in GTAP 4 to 15 sectors in GTAP 5. The trade and transport sector (t_t) in GTAP 4 was disaggregated into 4 sectors capturing trade (trd) and alternative modes of transportation including land (otp), water (wtp), and air (atp). Other private services (osp) in GTAP 4 was disaggregated into 6 sectors covering communications, financial services, insurance, business services, and recreational and other services.

The services trade data in based on IMF balance of payments statistics and was prepared with the help of the WTO. A new header, VTWR (m,i,r,s), representing the value of international margins by mode, is introduced in GTAP 5. The international margins data is based on data for U.S. that was contributed by Mark Gehlhar. The introduction of this new array necessitated the revision of the basic GTAP model to accommodate multiple modes of international transportation. The first pre-release of GTAP 5 does not include substantive bilateral content for services trade data yet. This should be included in the final release GTAP 5.

3.3 Energy Data

An upgrading of quality of the energy component of the GTAP data base was undertaken under a project sponsored by the U.S. Department of Energy. This resulted in GTAP-4E data base which is essentially the version 4 data base but with the energy section modified to be consistent with the independent price and volumes information which was collected from various sources. The key inputs to the energy data base are the International Energy Agency (IEA) volumes data and a supplementary energy price data base for 1995. This was done by upgrading the energy targets in the construction of the version 4 data base and then rerunning the FIT programs to hit the revised targets. Documentation on the GTAP-4E data base is available in Truong (2000). For the version 5 data base, 1997 volumes data were used and the 1995 energy prices data were simply updated to 1997 using price indices.

3.4 Other Data and Sources

Aside from the domestic data bases (input-output tables), and the services and energy data bases discussed above, construction of the database also requires other international data sets as well as data on factor splits and behavioral parameters. For the international data sets, data for 1997 were again obtained from the established data contributor as enumerated below:

(a) Macroeconomic Data. Data on GDP, population private consumption, government consumption, investment, and capital stock for 1997, all expressed in thousand US dollars, were obtained from the World Bank's database. For countries not covered by the World Bank data, GDP and population estimates were obtained from other sources such as the World Development Report and the CIA Fact Book.

(b) Input-Output Data on Agricultural and Food Products. 1997 data for GTAP 5 was again prepared by Everett Peterson of the VPI.

(c) Merchandise Trade Data. For GTAP 5, as well as with GTAP 3 and 4, merchandise trade data is supplied to GTAP by Mark Gehlhar of the ERS/USDA. The data for GTAP 5, updated to 1997, has improved coverage of some emerging economies and of Southern Africa.

(d) Protection Data. The primary source of data for merchandise tariffs is the UNCTAD TRAINS data set, data from which is supplied to GTAP by Jerzy Rozanzki and Emiko Fukase of the World Bank. For the first pre-release of version 5, coverage was limited to 44 standard country importers. We expect greater country coverage in subsequent pre-releases. Non-bilateralized tariff data for the Southern African countries, contributed by David Evans, is also used in GTAP 5. For agricultural protection data (production subsidies, export subsidies, and import tariffs) we use data on PSEs in OECD countries which were supplied to GTAP by Marinos Tsigas of the ERS/USDA. We expect significant changes in the protection data between now and the final release of GTAP 5.

(e) Factor Splits and Behavioral Parameters. Data on labor earnings by occupation and supply elasticities are used to generate the factor splits in the data base. Consumer demand parameters and other behavioral parameters are used in generating the parameters data file. No new data on factor splits and parameters were obtained for the first pre-release of GTAP 5. GTAP 4 data were simply expanded to agree with the version 5 regional and sectoral classification.

4 DATA BASE CONSTRUCTION PROCEDURE

The GTAP data base is composed of a sets file, a parameters file, and a data file. The sets file enumerates the regions, commodities, endowment factors, and other sets used in the GTAP model. The parameters file contains the behavioral parameters such Armington elasticities which are used by the GTAP model. The data file contains domestic and bilateral international trade flows of goods and services, measured in million US $, representing the state of the world economy for a particular benchmark year. The standard GTAP model can be used to operationalize the data base.

The broad procedure in the construction of the data base involves the preparation of global sets and mapping files used in the construction process; the simultaneous preparation of domestic data bases and international data sets; updating the regional data bases to match the macroeconomic, trade and protection data for the base year; and, final data assembly. In what follows, we provide a fuller description of the data construction procedure.

4.1 Preparation of Global Sets Files

The sets and mapping files that are used throughout the data base construction process are generated in a initial SETS module. These sets files include a list of regions, a list of sectors, a list of standard countries, a mapping file between the standard countries and the GTAP regions, and other files. The global sets file which includes the headers that are actually read in by the standard GTAP model is also created at the beginning of the process.

4.2 Domestic Data Tables

There are 55 primary regions in the first pre-release of GTAP 5. Primary regions are regions for which there is a contributed input-output table. These domestic data bases undergo a cleaning procedure and some of them also undergo a sectoral disaggregation procedure. Domestic data bases for the composite regions, as well as a representative input-output table, are also derived from the primary region tables.

Input-Output Tables: The requirements for contributing IO tables are specified in Huff, McDougall, and Walmsley (2000). The data contributor should ensure that the following are satisfied: (a) input-output structure requirements; (b) sectoral classification requirements - use GTAP sectoral classification or some aggregation thereof; (c) sign condition requirement - no negative flows, except for changes in stocks; (d) sectoral balance conditions: for each commodity, total sales should be equal to total costs; and (e) units: US $ million.

The mandatory commodity splits require the separation of agriculture and food processing, and energy, from other sectors. This supports the disaggregation procedure, allowing the use of special data sources for disaggregating agriculture and food processing. For GTAP 5, a package of software is made available to contributors of domestic data bases for their use in checking their data sets before submitting to Center staff.

Cleaning the IO tables: The contributed input-output tables which have satisfied the guidelines for contributors then go through a cleaning procedure at the GTAP center. Cleaning of input-output tables is done using from programs designed to check and ensure that the following conditions are met: (a) sign constraints, factor and commodity flows must be zero or positive; (b) international margins balance constraint, usage should equal supply; (c) income constraint, income by disposition should equal income by source; (d) import balance constraint, usage of imports should equal the supply of imports; and (e) domestic product balance, for each commodity, total usage less total cost should equal the production tax.

Representative Table : A representative table is created from the input-output tables which have full sectoral disaggregation. It is an input-output table created as a linear combination of these tables, using each region's GDP as weights. The representative table is used as a source of coefficients e..g. in the sectoral disaggregation procedure.

Disaggregation of IO tables : Greater disaggregation of food and agricultural sectors was introduced in GTAP 4. This posed a challenged specially since some of the input-output tables have only one aggregated agricultural sectors and one food processing sector while the GTAP 4 data base distinguished between 20 food and agricultural sectors (over the 11 in GTAP 3). In GTAP 5, as in GTAP 4, the sectoral disaggregation is accomplished using the maximum entropy method captured in specialized software developed at the Center and additional data from outside sources (Liu and McDougall, 1998). For food and agricultural sector disaggregation, Everett Peterson of VPISU, relied on the FAO supply utilization data base. He combined the FAO data with supplementary price information to create a country level data base of input-output data for each GTAP agricultural or food sector (Peterson, 1998).

For the disaggregation of the manufacturing sectors, especially where autos, parts, and electronic equipment had to be split in a given IO table, information from a representative table is used. The representative table is created from a simple summation of the set of IO tables for which full agricultural detail is available. Thus, for certain sectors which are not split in the original IO table of a given region, the structure of production, intermediate usage, and consumption for these sectors are adapted from the representative table (subject to control totals for the relevant cells within the aggregated transport sector).

Composite Region Tables: Composite regions are the GTAP regions for which there are no contributed tables. They are usually the "rest of" regions. The input-output tables for these regional groupings which account for the rest of the economies in the world that are not covered by the primary regions are constructed based on primary regions with similar GDP and economic characteristics as the countries that compose the composite regions. First, a mapping between the composite region member countries and the primary regions is constructed by finding the primary region with a comparable input-output structure to that of the composite region member country using as criteria GDP per capita, climatic similarity and intuition. Next, and input-output tables is created for each composite region member country using the input-output table of the corresponding primary region, rescaled to match GDP. Finally, the input-output table for each composite region is constructed by summing over the input-output tables of the its member countries.

4.3 International Data Sets

The international data sets include the macroeconomic data, trade data, protection data, and energy data. Raw or semi-processed data at the disaggregate (standard) country level are obtained from external data contributors. Further processing, including filling in of missing values and aggregation to the GTAP regional and sectoral classification are performed at the Center.

4.4 FITting the Data Sets

Since the reference periods of the input-output tables vary, the domestic data bases have to be updated for the base year of the GTAP data base (1997 for GTAP 5). The general approach is to update these tables by reconciling them with macroeconomic data for the base year and then combining the updated domestic data bases with the international data sets using the general reconciliation procedure called FIT developed at the Australian Industry Commission (James and McDougall, 1993).

FIT is an economic model of a regional economy that allows targeting of economic aggregates to match conditions in a particular year. Shocks are applied to the target variables and changes to all other variables in the model in response to the shocks are computed. FIT maintains market clearing and zero profit conditions while keeping primary factor prices fixed and allowing value-added and intensity of input usage to adjust, subject to a pre-specified penalty function. The variables that are targeted in updating each region's input-output table are: aggregate household consumption expenditures, aggregate government spending, aggregate expenditures for gross capital formation, exports by commodity, imports by commodity, export subsidies, import tariffs, production subsidies/taxes, and selected import prices of energy commodities. Using FIT to update an input-output table with macroeconomic, trade, protection and energy information, as described above, results in an internally consistent regional data base. The regional data bases are put together to construct an interim global data file.

4.5 Parameter Files and Factor Splits

The global parameters file is created somewhat independently from the global data file. The expansion and substitution parameters for the CDE utility function are calibrated using some initial deriveprice and income elasticity targets using a maximum entrophy procedure which is specified in GAMS. The region-generic Armington trade elasticities and the elasticities of substitution in value-added are those from the SALTER project, mapped to the GTAP sectoral classification.

For the primary factor shares in agriculture, external estimates of factor earnings share

are applied to agricultural value-added. For primary factor shares in the natural resource based sectors, a proportion of the earnings of labor and capital are reallocated to natural resources to achieve target supply elasticities. This is done by determining a share which when combined with the elasticities of substitution in the model, replicates a target level of supply response, based on estimates in the literature. In splitting the labor endowment into skilled and unskilled components for each region and sector, the labor payments are disaggregated using labor earnings shares from the labor data set.

4.6 Data Assembly

The interim global data base that is created from the updated domestic data bases is subjected to checks for global consistency and balance. Supplementary information on primary factor splits in agriculture, as well data on capital stock and depreciation, are brought into the data base. The bilateral trade detail from the trade matrix, the bilateral detail in the protection data, and the international trade and transport margins are also incorporated into the data base at this point. The result is a global GTAP data base.

5 DATA BASE DEVELOPMENT

One of the more important innovations related to GTAP 5 is the way in which it is being produced. Relative to how earlier versions of the data base were produced, the data base construction procedure has seen immense improvement such that it is now completely automated, replicable, and flexible. This section provides a discussion of the innovations that have been introduced in the GTAP 5 data base construction process to better handle the large and complex task.

5.1 Modular Structure

We implement the construction process, described in the previous section, as a collection of modules arranged in a directory tree.

The data base construction procedure is divided into several modules. For example, there is a trade module where merchandise and services trade data is prepared. There is a module wherein IO tables for the composite regions are created. There is also a FIT module wherein the input/output table for each region is reconciled with the trade, protection, macroeconomic, and energy data. Each module may be run separately or we can run all modules collectively, under a master program.

The data base construction root directory contains the master make description files, various top-level module directories, and a module-generic directory. The root directory contains no data or program files (other than the make description file); these are all held in subdirectories.

Modules may contain submodules; those that do are laid out similarly to the construction root directory, with all data and program files relegated to module-specific subdirectories and a module-generic subdirectory.

Within each bottom-level module we create a standard module directory structure. This structure collects files according to their function and treatment within the module. The module root directory contains as many as needed of the following files and directories (directories are marked by an appended slash character /):

Makefile : make description file

lcl/ : data input files which are local or specific to the module

in/ : Data input, including set and mapping information.

src/ : Program source. This includes `TABLO' source code, stored input ("sti") files, and command files. It includes also any scripts and batch files used in the module.

wrk/ : Working files. We include here files that play a purely intermediate role within the module; that is, they are created within the module, and they are used in creating other files within the module but they are not amongst the final outputs or reports. They include both intermediate data files, but also generated program files, such as Fortran files and executable files generated from `TABLO' source code.

dmp/ : Dumpable reports. We put here log files and the like, that may provide useful information when things go wrong, but may safely be discarded after a successful run.

out/ : Data output. Files in this directory should be either final data base construction files, for inclusion in the GTAP data distribution, or files required for use in other modules.

A module may also have the following files or directories :

index : List of top-level files and directories. Standard files and directories may be omitted. If there are no non-standard files or directories, the entire `INDEX' file may be omitted.

readme : Introductory information for module users. It may refer the reader to other files, an `INDEX' file or detailed documentation in the `doc' directory. If there is nothing special to be said about the module, the entire `README' file may be omitted.

pre/ : If the module involves some initial formatting of source data files, not under build management, this is where we keep the relevant files.

doc/ : Documentation.

rep/ : Reports. We include here files we wish to preserve, for information of permanent value about data construction, but which do not form part of the data package and are not needed for later processing.

5.2 Build Management

We automate the construction process using the program make. The make program identifies the command needed to create or update the data base, runs those commands, and displays an error message and stops if it encounters an error in running them.

To tell make how to create the data base, we maintain a makefile showing which target and intermediate files depend on which intermediate and original files, and the commands needed to create the targets files from the files on which they depend. The makefile also contains brief comments, providing basic documentation of the build procedure.

To keep the makefile manageable, to support a modular structure, we use make recursively. The master makefile does not contain any of the commands directly used to create the data base; instead it shows how the different modules depends on each other – how each modules outputs are another modules inputs – and contains commands to run make on each module. These recursive make commands read information from module-specific makefiles.

In practice, we do not bring the entire construction process under build management. More specifically, we do not bring under build management some procedures involved in the initial conversion or formatting of original data files. We do however automate these steps as far as practicable.

We confine these unmanaged procedures to initial file conversion and formatting. Their function is to convert the data to GEMPACK-readable form, or some other form which we can handle automatically, as directly as possible. All substantive data processing, as opposed to formatting, is done under build management.

5.3 Version Control

Version control systems have now been implemented in the data construction process. A publicly available software, Concurrent Version Systems (CVS), developed by Cyclic Software, is used with the program files. Under this system, each program file is stored in a repository. Modifications to each file, corresponding to a version of the file, is also stored. Each version is tagged and can be retrieved. The use of CVS enables more efficient use of computer disk space since only the latest version of each program file is actually stored as opposed to completely storing all versions of a file.

The Data Version System (DVS), developed by Robert McDougall at the Center, is used for storing input data files (including header array files) since CVS does not handle HAR files very well.

The version control systems are very important for quality control in the complex data base construction procedure. It enables the developer to start from a clean build each time, i.e. start with an empty directory structure and populate the directory, in automated fashion, with the latest program code files and also with the latest version of the input files. The data developer can also readily recreate a previous version of an output data file or rebuild an earlier version of the database by selecting from the version archive system the appropriate set of files used to create that version.

5.4 Flexibility

The data base construction process should lend itself to the development of new revisions or special versions. In particular, it should be easy to use revised versions of original input data sets supplied by external contributors and to change the regional classifications.

There are two main aspects to flexibility. We want to make it easy to revise the data base when we get new source data. And we want to be able easily to revise the regional classification.

To make it easy to revise the data base when we get new source data, we keep data and programs in separate files. We count as data not only economic flow data but also set and mapping information. To make it easy to revise the regional classification, we rely partly on the practice described above, but also on another practice, encouraging contributors of multi-country data sets to provide the data on a country basis rather than on a GTAP region basis. This lets us revise the GTAP regional classification without going back to data contributors for new reclassified source data.

Regional flexibility is achieved when new regions can be added by simply including the new region(s) in a list or regions used in the data construction process and supplying the additional country's input output table.

In previous versions of the database, most source international data was made available to data developers at the aggregated GTAP region level. Starting in version 4, international data bases were collected at the country level. The data is then mapped or transformed to apply to a standard set of countries. A mapping between the standard set of countries and the set of regions is then used to aggregate the data. This mapping file should be revised when a new region is introduced in the GTAP data base. The used of standard country data bases is useful in achieving regional flexibility. Regional flexibility is also facilitated by ensuring that there is no hard-coding of regions in the programs codes. The set of region and mappings should be read from files generic to the construction process.

5.5 Common Tool Set

The data base construction process should use a common tool set, to spare developers the need to search for and reimplement tools already on hand, learn multiple versions of the same tool, or maintain multiple version of the same tool; and to make it easy to use the same tool set as originally used when replication old builds.

For data processing, the tool of choice is GEMPACK software suite, which is also used to implement the standard GTAP model. This is because most of the existing programs in the data base construction package use GEMPACK. The more we stick with GEMPACK, the less time we waste in converting file formats, for instance between GEMPACK and `GAMS'.

For build management we use MAKE.

For formatting and text processing, we use SED, AWK, and PERL since these are all available as free software in implementations of excellent quality. The programs AWK and PERL overlap in function. AWK is simpler and easier to learn, but PERL is now more widely known.

For batch jobs, where these are not conveniently handled with DOS batch files within COMMAND.COM, we use BASH and shell scripts written for BASH. However there should be little need for this; for the most part, we can run the necessary command sequences direct from the make file.

6 CONCLUDING REMARKS

A publicly available, fully-documented, global data base is the centerpiece of the Global Trade Analysis Project. Since GTAP's inception in 1993, the data base, along with the other components of the GTAP have continued to evolve and grow in response to and with the support of the users of the data base. Over the last seven years, the GTAP data base has supported quantitative economic analysis using the GTAP model and other multi-regional, applied general equilibrium models. As the regional and sectoral classification expands and as new features are added to the data base, the size and complexity of the data base construction task also increases. The adaption of new methods and standards such as the use of a modular structure, build management using a recursive MAKE structure, version control for the program files and input data files, regional flexibility, and use of a common tool set, has made complete automation, replicability and regional flexibility in the construction of the GTAP 5 (pre-release 1) data base.

The innovations in the data base construction process should greatly facilitate the construction of a complete version of the data base after considerable changes have been incorporated, thus allowing for not just one pre-release but for several pre-releases before the final GTAP 5 data base. The changes which are likely to occur between now and final GTAP 5 include the updating of some domestic data bases and the addition of a few more regions such as Egypt and Tunisia. Better regional coverage and better quality of the protection data, particularly the merchandise tariffs and the agricultural protection measures are also planned. More substantive detail in the bilateral services trade data is also forthcoming.

7 REFERENCES

Dimaranan, Betina, Thomas Hertel, and Robert McDougall. 1997. "Aggregation and Calibration," Chapter 20 in Global Trade, Assistance, and Protection: The GTAP 3 Data Base, Robert A. McDougall (editor).

Dimaranan, Betina and Robert McDougall. "Behavioral Parameters," Chapter 18 in Global Trade, Assistance, and Protection: The GTAP 4 Data Base, Robert A. McDougall (editor).

Dimaranan, Betina, Jing Liu, Robert McDougall, Thomas Hertel, and Yves Surry. 1998. "CDE Calibration," Chapter 21 in Global Trade, Assistance, and Protection: The GTAP 4 Data Base, Robert A. McDougall, Aziz Elbehri and Truong P. Truong (editors).

Gehlhar, Mark. 1997. "Reconciling Bilateral Trade Data for Use in GTAP," Chapter 11 in Global Trade, Assistance, and Protection: The GTAP 3 Data Base, Robert A. McDougall (editor).

Gehlhar, Mark. 1998. "Reconciling Bilateral Trade Data for Use in GTAP," Chapter 11 in Global Trade, Assistance, and Protection: The GTAP 4 Data Base, Robert A. McDougall, Aziz Elbehri and Truong P. Truong (editors).

Hertel, Thomas W. (editor). 1997. Global Trade Analysis: Modeling and Applications. New York: Cambridge University Press.

Hertel, Thomas W. 2000. "The Global Trade Analysis Project: Issues and Future Directions," Background Paper for the GTAP Advisory Board Meeting, Purdue University, 12-14 April.

Hertel, Thomas W. 1998. "Introduction," Chapter 1 in Global Trade, Assistance, and Protection: The GTAP 4 Data Base, Robert A. McDougall, Aziz Elbehri and Truong P. Truong (editors).

Hertel, Thomas and Marinos Tsigas. 1998. "Primary Factor Shares and Supply Response in Natural-Resource Based Industries," Section 17.4 in Global Trade, Assistance, and Protection: The GTAP 4 Data Base, Robert A. McDougall, Aziz Elbehri and Truong P.

Truong (editors).

Huff, Karen, Robert McDougall, and Terrie Walmsley. 1999. “Contributing Input-Output Tables to the GTAP Data Base,” GTAP Technical Paper No.1, Release 4.1, August.

Ingco, Merlinda. 1997. “Agricultural Protection,” Chapter 14 in Global Trade, Assistance, and Protection: The GTAP 3 Data Base, Robert A. McDougall (editor).

James, M. and Robert A. McDougall, 1993. “FIT: An Input-Output Data Update Facility for SALTER,” SALTER Working Paper No.17, Canberra, Australia: Australian Industry Commission.

Liu, Jing and Robert A. McDougall. 1998. “Disaggregation of Input Output Tables,” Chapter 16 in Global Trade, Assistance, and Protection: The GTAP 4 Data Base, Robert A. McDougall, Aziz Elbehri and Truong P. Truong (editors).

Liu, Jing, Nico van Leeuwen. Tri Thanh Vo, Rod Tyers, and Thomas Hertel. “Disaggregating Labor Payments by Skill Level,” GTAP Technical Paper No. ?.

Liu, Jing, Yves Surry, Betina Dimaranan, and Thomas Hertel. 1998. “CDE Calibration,” Chapter 21 in Global Trade, Assistance, and Protection: The GTAP 4 Data Base, Robert A. McDougall, Aziz Elbehri and Truong P. Truong (editors).

McDougall, Robert A. (editor). 1997. Global Trade, Assistance, and Protection: The GTAP 3 Data Base. Center for Global Trade Analysis, Purdue University.

McDougall, Robert A., Aziz Elbehri and Truong P. Truong (editors). 1998. Global Trade, Assistance, and Protection: The GTAP 4 Data Base. Center for Global Trade Analysis, Purdue University.

Peterson, Everett. 1998. “Disaggregating Agricultural and Food Sectors,” Chapter 15 in Global Trade, Assistance, and Protection: The GTAP 4 Data Base, Robert A. McDougall, Aziz Elbehri and Truong P. Truong (editors).

Reincke, Ulrich. 1997. “Import Tariffs on Merchandise Trade,” Chapter 13 in Global Trade, Assistance, and Protection: The GTAP 3 Data Base, Robert A. McDougall (editor).

Tsigas, Marinos and Thomas W. Hertel. 1997. “Primary Factor Shares in Agriculture,” Section 17.3 in Global Trade, Assistance, and Protection: The GTAP 3 Data Base, Robert A. McDougall (editor).

Table 1. Sectoral Classification, GTAP Data Bases 2, 3, and 4

GTAP 1 - 3		GTAP 4	
pdr	Paddy rice	pdr	Paddy rice
wht	Wheat	wht	Wheat
gro	Cereal grains n.e.c	gro	Cereal grains n.e.c
ngc	Non-grain crops	v_f	Vegetables, fruits, nuts
wol	Wool	osd	Oil seeds
olp	Other livestock	c_b	Sugar cane, Sugar beet
for	Forestry	pfb	Plant-based fibers
fsh	Fisheries	ocr	Crops nec
col	Coal	ctl	Cattle, sheep, etc.
oil	Oil	oap	Animal products nec
gas	Gas	rmk	Raw milk
omn	Other minerals	wol	Wool, silk-worm cocoons
pcr	Processed rice	for	Forestry
met	Meat products	fsh	Fishing
mil	Milk products	col	Coal
ofp	Other food products	oil	Oil
b_t	Beverages and tobacco	gas	Gas
tex	Textiles	omn	Minerals nec
wap	Wearing apparel	cmt	Cattle, sheep, etc. meat products
lea	Leather etc	omt	Meat products nec
lum	Lumber	vol	Vegetable oils and fats
ppp	Pulp, paper, etc	mil	Dairy products
p_c	Petroleum and coal	pcr	Processed rice
crp	Chemicals, rubbers, and plastics	sgr	Sugar
nmm	Non-metallic minerals	ofd	Food products nec
i_s	Primary ferrous metals	b_t	Beverages, tobacco products
nfm	Nonferrous metals	tex	Textiles
fmp	Fabricated metal products	wap	Wearing apparel
trn	Transport industries	lea	Leather products
ome	Machinery and equipment	lum	Wood products
omf	Other manufaturing	ppp	Paper products, publishing
egw	Electricity, water and gas	p_c	Petroleum, coal products
cns	Construction	crp	Chemical, rubber, plastic products
t_t	Trade and transport	nmm	Mineral products nec
osp	Other services (private)	i_s	Ferrous metals
osg	Other services (government)	nfm	Metals nec
dwe	Ownership of dwellings	fmp	Metal products
		mvh	Motor vehicles, parts
		otn	Transport equipment nec
		ele	Electronic equipment
		ome	Machinery, equipment nec
		omf	Manufactures nec
		ely	Electricity
		gdt	Gas manufacture, distribution
		wtr	Water
		cns	Construction
		t_t	Trade, transport
		osp	Financial, business, etc. services
		osg	Public admin., education, etc.
		dwe	Dwellings

Table 2. Regional Classification, GTAP Data Bases 2, 3 and 4

GTAP 2		GTAP 3		GTAP 4	
aus	Australia	aus	Australia	aus	Australia
nzl	New Zealand	nzl	New Zealand	nzl	New Zealand
can	Canada	jpn	Japan	jpn	Japan
usa	United States	kor	South Korea	kor	South Korea
jpn	Japan	idn	Indonesia	idn	Indonesia
kor	South Korea	mys	Malaysia	mys	Malaysia
e_u	European Union	phl	Philippines	phl	Philippines
idn	Indonesia	sgp	Singapore	sgp	Singapore
mys	Malaysia	tha	Thailand	tha	Thailand
phl	Philippines	chn	China	vnm	Vietnam
sgp	Singapore	hkg	Hongkong	chn	China
tha	Thailand	twn	Taiwan	hkg	Hongkong
chn	China	idi	India	twn	Taiwan
hkg	Hongkong	sas	Rest of South Asia	ind	India
twn	Taiwan	can	Canada	lka	Sri Lanka
arg	Argentina	usa	United States	ras	Rest of South Asia
bra	Brazil	mex	Mexico	can	Canada
mex	Mexico	cam	Central America, Caribbean	usa	United States
lam	Rest of Latin America	arg	Argentina	mex	Mexico
ssa	Sub-Saharan Africa	bra	Brazil	cam	Central America, Caribbean
mna	Mid. East and N. Africa	chl	Chile	ven	Venezuela
eit	Economies in Transition	rsm	Rest of South America	col	Colombia
sas	South Asia	e_u	European Union 12	rap	Rest of Andean Pact
row	Rest of World	eu3	Austria, Finland, and Sweden	arg	Argentina
		eft	European Free Trade Area	bra	Brazil
		cea	Central European Associates	chl	Chile
		fsu	Former Soviet Union	ury	Uruguay
		mna	Middle East and North Africa	rsm	Rest of South America
		ssa	Sub-Saharan Africa	gbr	United Kingdom
		row	Rest of World	deu	Germany
				dnk	Denmark
				swe	Sweden
				fin	Finland
				reu	Rest of European Union
				eft	EFTA
				cea	Central European Associates
				fsu	Former Soviet Union
				tur	Turkey
				rme	Rest of Middle East
				mar	Morocco
				rnf	Rest of North Africa
				saf	South African Customs Union
				rsa	Rest of Southern Africa
				rss	Rest of Sub-Saharan Africa
				row	Rest of World

Table 3. List of Regions and Sectors in GTAP 5 pre-release 1

GTAP 5 pre-release 1 regions					
aus	Australia	per	Peru	che	Switzerland
nzl	New Zealand	ven	Venezuela	xef	Rest of EFTA
chn	China	xap	Rest of Andean Pact	hun	Hungary
hkg	Hongkong	arg	Argentina	pol	Poland
jpn	Japan	bra	Brazil	xce	Central European Associates
kor	South Korea	chl	Chile	xsu	Former Soviet Union
twn	Taiwan	ury	Uruguay	tur	Turkey
idn	Indonesia	xsm	Rest of South America	xme	Rest of Middle East
mys	Malaysia	aut	Austria	mar	Morocco
phl	Philippines	dnk	Denmark	xnf	Rest of North Africa
sgp	Singapore	fin	Finland	bwa	Botswana
tha	Thailand	fra	France	xsc	South African Customs Union
vnm	Vietnam	deu	Germany	mwi	Malawi
bgd	Bangladesh	gbr	United Kingdom	moz	Mozambique
ind	India	grc	Greece	tza	Tanzania
lka	Sri Lanka	irl	Ireland	zmb	Zambia
xsa	Rest of South Asia	ita	Italy	zwe	Zimbabwe
can	Canada	nld	Netherlands	xsf	Rest of Southern Africa
usa	United States	prt	Portugal	uga	Uganda
mex	Mexico	esp	Spain	xss	Rest of Sub-Saharan Africa
xcm	Central America, Caribbean	swe	Sweden	xrw	Rest of World
col	Colombia	xbl	Belgium/Luxembourg		

GTAP 5 pre-release 1 sectors					
pdr	Paddy rice	omt	Meat products nec	otn	Transport equipment nec
wht	Wheat	vol	Vegetable oils and fats	ele	Electronic equipment
gro	Cereal grains n.e.c	mil	Dairy products	ome	Machinery, equipment nec
v_f	Vegetables, fruit, nuts	pcr	Processed rice	omf	Manufactures nec
osd	Oil seeds	sgr	Sugar	ely	Electricity
c_b	Sugar cane, Sugar beet	ofd	Food products nec	gdt	Gas manuf., distribution
pfb	Plant-based fibers	b_t	Beverages, tobacco products	wtr	Water
ocr	Crops nec	tex	Textiles	cns	Construction
ctl	Cattle, sheep, etc.	wap	Wearing apparel	trd	Trade
oap	Animal products nec	lea	Leather products	otp	Transport nec
rmk	Raw milk	lum	Wood products	wtp	Water transport
wol	Wool, silk-worm cocoons	ppp	Paper products, publishing	atp	Air transport
for	Forestry	p_c	Petroleum, coal products	cmn	communication
fsh	Fishing	crp	Chem., rubber, plastic prods	ofi	Financial services nec
col	Coal	nmm	Mineral products nec	isr	Insurance
oil	Oil	i_s	Ferrous metals	obs	Business services nec
gas	Gas	nfm	Metals nec	ros	Recreational, other services
omn	Minerals nec	fmp	Metal products	osg	Public adm., defense, etc.
cmt	Cattle, sheep, etc meat prods	mvh	Motor vehicles and parts	dwe	Dwellings