International Services Trade Data for CGE Modellers,
or,
Entropy Methods for Data Reduction

Robert A. McDougall

2000-06

Abstract

This paper reports progress in the construction of a services trade data set for use by general equilibrium modelers, and in constructing the GTAP data base. It describes briefly the background and aims of the services trade data project, the contents of the services trade data prepared for the current GTAP prerelease, and the data construction process. It also illustrates the use of entropy-theoretic methods in data investigation, to direct search for errors in and artifacts of the construction process, and to highlight key features of the data.

This paper has not both a subject, the GTAP services trade data project, and a theme, the use of entropy methods in describing and reviewing data.

The services trade data project undertakes to develop a services trade data set to support services trade policy analysis, and for incorporation in the GTAP data base. Project deliverables include this paper, a revised GTAP model, a revised GTAP data structure, a revised GTAP sectoral classification, procurement of new source I/O tables, a new trade in services data set, a revised trade module, a revised protection module, and a revised data base aggregation program. Preliminary versions of all these have been provided in the GTAP 5 prerelease. In this paper we discuss only the services trade data.

1 Background

The Global Trade Analysis Project (*GTAP*) has as one of its main activities the development and maintenance of a multi-country database for computable general equilibrium (*CGE*) modeling. The database has gone through several revisions, the latest publicly released revision being GTAP 4 (McDougall, Elbehri and Truong [11]).

The database is assembled at the Center for Global Trade Analysis, Purdue University, using data contributed by GTAP Consortium members and other researchers around the world.

A key part of the GTAP database is the trade data arrays. These contain data on the value of trade, by commodity, region of origin, and region of destination. There is also a smaller array showing the supply of international margin services, that is, freight transport services used in international trade.

For merchandise trade data we have a good single source, a dataset maintained by Mark Gehlhar at the Economic Research Service, United States Department of Agriculture ([3]). For these data, we at the Center have had little more to do than to aggregate it to the current GTAP regional classification and copy it into our trade data set.

For services trade data we have had to become more deeply involved. Here we have had no single satisfactory source of data. We have constructed the data ourselves using several sources:

- for total services exports and imports by region, IMF balance of payments statistics;
- for the sectoral composition of services exports and imports in each region, the region's input-output table, and
- for the bilateral trade pattern, a trade data set originally prepared for the Michigan world CGE model.

We have combined these various data sources and reconciled their inconsistencies to construct a GTAP services trade dataset, which we then incorporate first into a comprehensive trade dataset and then into the GTAP database.

This approach has a number of limitations.

- For some regions we have no contributed input-output table. For these regions we synthesize a table in-house. The resultant table should meet minimal sanity checks, but cannot be considered a true data source for the region. Yet this synthesized table is the source we use for the sectoral composition of services exports and imports.
- Even for regions for which we have contributed tables, the tables relate to different reference years and use a variety of conventions and sectoral classifications. They do not together form a consistent data source for services trade, comparable to the merchandise trade data set.
- The source for the bilateral trade pattern, the Michigan model data set, has several limitations.
  - It is out of date (its reference year is 1990).
  - It is based on a variety of incomplete and inconsistent data sets, reflecting the poor availability of services trade source data at the time of construction.

- The sectoral and regional classifications are inconsistent with and more highly aggregated than the GTAP classifications.
- As we have found, we need to distort it quite severely to fit it to the other services trade data.
- It is no longer being maintained.

At the same time, there is an increasing demand for accurate and detailed services trade data. This reflects the increasing importance of trade in services, the increasing emphasis on services in international trade negotiations, and movements in some countries towards domestic deregulation and privatisation of some service industries. Several GTAP Consortium members have indicated an interest in services trade policy analysis, directed in part toward future rounds of negotiation within the General Agreement on Trade in Services (*GATS*).

From this perspective, the representation of trade in services in the GTAP data base has several limitations. The classification of services sectors is not detailed, and differs from that used in the GATS. There is a lack of data on barriers to trade and other forms of services industry assistance. And some of the delivery modes identified in the GATS do not appear in the data base at all.

While statistics on trade on services remain less than complete, there have been major improvements recently in the international statistics compiled by the International Monetary Fund and by Eurostat.

2 The services trade project

Between the problems with the old data set, the increasing demands of CGE modelers, and the opportunities created by improvements in source data, it is timely to upgrade the treatment of services in the GTAP data base. The services trade project aims to develop a services trade data set that:

- draws on the new services trade statistics available from the IMF and EUROSTAT,
- can be maintained on an ongoing basis,
- has its own intrinsic interest and usefulness, and
- can serve as the basis for the services trade data in the GTAP data base.

The services trade project covers procurement and preparation of source data sets, based on IMF and EUROSTAT services trade data, by the World Trade Organization, and by Wusheng Yu at Purdue (Luanga and Yu [10]), construction of the services trade data set, and revision of the GTAP trade data modules to use the new data set.

The services trade data set:

- contains data not for GTAP regions but for individual countries,
- uses for its sectoral classification a refinement of the GTAP sectoral classification revision 1 (*GSC1*, the sectoral classification used in GTAP 4, as it pertains to services,
- may contain (if demand is apparent) some features not required to support the GTAP data base, for example,
  - separate data for the GATS supply mode *consumption abroad* from the modes *cross-border supply* and *temporary entry of natural persons*,
  - data for royalties and licensing fees, and
  - bilateral trade data for margin services.

On the negative side, the project does not aim to provide:

- services protection data, or
- data not just for trade as conventionally defined but for all GATS supply modes, in particular, *commercial presence,*

though we do not rule out enhancements in these directions in future versions.

As we pursue the services trade project, we are also working on the next version of the GTAP data base, GTAP 5. GTAP 5 relies on the services trade project for its services trade data source.

The state of play with GTAP 5 and the services trade data project is:

- We have prepared a first interim version of the services trade data set.
- We have used that services trade data set in the first GTAP 5 prerelease (available to Consortium members only).
- We have made some bugfixes to the services trade data set, reflected in this document but not in the GTAP 5 prerelease.
- The current version of the services trade data set draws on IMF data for services exports and imports for 1997. It does not yet utilize IMF data for other reference years, nor the EUROSTAT bilateral trade data.

3 Services trade data in the GTAP 5 database

The main purpose of the services trade data set is of course to improve the data for service commodities in the trade arrays in the GTAP data base. The GTAP 4 data base included trade data for 45 regions and 8 service commodities. This is however slightly misleading,

Table 1: Service sector breakout between GTAP 4 and 5

old	new
electricity	electricity
gas distribution and transmission	gas distribution and transmission
water	water
construction	construction
trade and transport	trade
	transport n.e.c.
	water transport
	air transport
	communication
financial, business, recreational services	financial services n.e.c.
	insurance
	business services n.e.c.
	recreational, other services
public administration, defence, education, health	public administration, defence, education, health
ownership of dwellings	ownership of dwellings

since of the 8 service commodities, only 5 are traded internationally. Gas distribution, water, and ownership of dwellings are not traded to any great extent; in the GTAP data base we represent them as not traded at all.

In GTAP 5 we increase the number of regions to 65, and the number of service commodities to 15. Again, 3 service commodities are essentially non-traded, so we have trade data for 12.

Table 1 shows how we have broken out the service sectors between GTAP 4 and GTAP 5. It may be observed that the new sectors are concentrated in transport and business services.

The new sectoral classification has implications for the treatment of internation margins. In each previous GTAP database, there has been just one margin commodity, trade and transport. Accordingly, we have not stored margin usage data explicitly in the database, but have calculated margin usage for each trade flow as the difference between that flow's CIF and FOB valuations. Now however we have several margin commodities, water transport, air transport, and "other" transport. Now therefore we need to store margin usage data specifically, to show how the CIF-FOB value gap is composed of air, water, and "other" transport margins. As before, we have also an array showing margin services supply for each of the three margin commodities and for each region.

In principle, we could replace the margin usage and margin supply arrays with a single fifth-order array, showing international margins by (1) margin commodity, (2) region of

Table 2: Arrays with services trade data in GTAP 5

name	dimensions	description
VXMD	(TRAD,REG,REG)	non-margin exports of %1 from %2 into %3, at %3 market prices
VXWD	(TRAD,REG,REG)	non-margin exports of %1 from %2 into %3, at FOB prices
VMWS	(TRAD,REG,REG)	imports of %1 from %2 into %3, at CIF prices
VMWS	(TRAD,REG,REG)	imports of %1 from %2 into %3, at %3 market prices
VST	(MARG,REG)	margin exports of %1 from %2
VTWR	(MARG,TRAD,REG,REG)	margin usage of %1 in facilitation of flow of %2 from %3 to %4

MARG margin commodity
TRAD tradable commodity
REG region
%*n* argument *n* from dimension list

origin of the margin, (3) commodity traded (i.e., the commodity whose international flow is facilitated), (4) region of origin of the commodity traded, and (5) destination region. We do not adopt this course, largely because the resulting very large array would be liable to run up against memory constraints, reducing the dimensions of the database that could be used in simulations or even viewed on any given machine.

We could on the other hand store no margin usage array, but instead assume that the composition of margins is the same for all transported commodities and all routes ((origin, destination) pairs). We do not adopt that course, because it would imply that for example coal and electronic equipment have the same water:air transport shares, and that goods sent to Japan and Switzerland have the same air:land transport shares ("other" transport comprising mainly road and rail transport). Rather than accept that, it is both desirable and practicable to provide information on how shares vary across commodities and routes.

Table 2 shows arrays containing services trade data in the GTAP 5 data base. The trade data base includes four arrays for non-margin trade flows. For services, the two arrays showing values at world prices, the FOB and CIF value arrays, contain equivalent data, since we do not recognize any margins on services. The arrays for exports and imports at market prices contain little new data; we have little information on services border protection, so the values at market prices differ little from the values at world prices. So essentially for non-margin services flows we have one array of data for 12 commodities, 65 source regions, and 65 destinations, or 50,700 elements in total.

For margin usage, we have non-zero usage only for 41 merchandise trade commodities.

So we have data for 3 margin commodities, by 41 transported commodities, by 65 regions of origin, by 65 destinations, or 519,675 elements. For margin supply, we have 3 by 65 or 195 elements. Overall then we have about six hundred thousand services trade data values, most of which pertain to margin usage.

4 Entropy theory and data reduction

Generating a data set of almost six hundred thousand elements is in itself a success at some level, but one that brings its own problems; in particular, the problem of how such a data set can be inspected. We wish to inspect the data firstly, because we consider them of some intrinsic interest apart from their utility in supporting CGE simulations, secondly, to understand how they differ from other related data sets, including but not limited to earlier GTAP databases, and thirdly, to search for possible errors and misfeatures in the data construction process.

To make this humanly possible we need methods for data reduction—for converting small data sets to large ones, that retain as much as possible of the key information from the large data sets. Since each method provides only a partial view, it is likely that we will want to adopt several methods, each helping to supply the deficiencies of the others. One familiar method is calculating marginal subtotals—for example, from a third order bilateral trade array, trade by commodity, region of origin, and destination calculating subtotals, first order arrays for trade by commodity, or by region of origin, or of destination. Another is aggregation, reducing for example the 65 regions of GTAP 5, or the 12 main service commodities, to some smaller number.

Each of these methods *condenses* a data set into a smaller set that covers the same material but at a lower level of detail. The smaller data set shares no common elements with the larger, but it typically conserve some key properties; in the cases we have described, it conserves for example the array totals. We shall examine here a different approach, that *selects* from the original data set elements that are of special significance, or especially surprising, or that carry a high information content. The elements of the new data set are a selected subset of the old data set, but have no significant overall properties (for example, their totals are of no particular significance). In this approach, we may for example select the data elements that differ most from earlier versions of the same data set, or from a proxy data set, or that are least predictable from a high level or condensed view of the current data set.

To develop a data selection approach, we need to operationalise the vague notion alluded to above, a measure of the *information content* of data elements. The idea is to assign an information measure to each element, sort the elements by this measure, and select the most informative. Since we want to make the selection against some background knowledge, for example, an older version of the same data set, or a proxy data set, we need a relative measure of information content, information relative to some other data set. We find the measures we need in the field known (not coincidentally) as information theory, and more especially in the subfield known as entropy theory.

Information theory has its origins in the early decades of the twentieth century, but received its first major impetus from Shannon [13]. Shannon's interest was in communication engineering, but connections were established to other fields such as computer science (Kolmogorov complexity, [8] 1968), finance theory (log optimal portfolios, Kelly [7] 1956), spectral analysis in geophysics and elsewhere (Burg [2] 1975), and statistics (contingency tables, Fisher's information measure; see, e.g., Kullback [9]). Shannon appropriated from statistical mechanics the term *entropy* for one of the key concepts of the theory; the analogy implicit in the term has been vigorously expounded by Jaynes (e.g., [6]).

Theil [14] and Bacharach [1] pioneered the application of entropy theory to problems of economic data construction. Though Theil discussed a variety of applications, subsequent interest focussed mostly on matrix balancing applications (for an overview, see Schneider and Zenios [12]). More recently, Judge and associates have developed and popularized applications of entropy-theoretic methods to a wider range of data construction and estimation problems (see especially Golan, Judge and Robinson [5] and Golan, Judge and Miller [4]).

Further information on the history of the RAS may be obtained from Bacharach [1] and Schneider and Zenios [12].

From entropy theory, we have a number of concepts that deal with distributions of a given total, or sets of shares that sum identically to one. In particular, we have the concept of *information gain* or *cross entropy*, a measure of the information contained in an observed distribution, relative to a previous estimate of the distribution. If the observed distribution is equal to the estimated distribution, the information gain or cross entropy is zero; otherwise, the cross entropy is strictly positive. The cross entropy tends to be large when items that had a very small share in the estimated distribution have a large share in the observed distribution.

With estimated shares s_i^0 and observed shares s_i^1, the cross entropy is

$$\sum_i s_i^1 \log \frac{s_i^1}{s_i^0}.$$

The base of logarithms is a matter of taste. In this paper we use logarithms to base 2. Entropies calculated in this way are said to be measured in *bits*, since by this measure one binary digit has an information content of 1.

For the data selection task, we want to identify the elements i of the data array that contribute the most relative information, or contain the most significant differences against the comparison data array. How might we operationalise these vague notions? There is surely no unique solution to this problem; yet we can stipulate some properties we should seek in any proposed solution.

Consider this problem: comparing export levels against GDP. We have data for exports by region, and cursory inspection shows that large regions tend to have large export values.

We want to progress beyond this near-truism, and identify the aspects of the data set that stand out against it: the large regions with relatively small export values, or the small regions with relatively high export values.

We would like to have a measure of the difference between the two sets of shares, that can be decomposed into region-wise components, such that:

1. The greater the absolute difference, the greater the significance. A billion dollar difference is more significant *cet. par.* than a million dollar difference.

2. The greater the relative difference, the greater the significance. Between two billion dollar differences, if one represented if one represented 200 per cent of the initial total and the other 50 per cent, we would regard the 200 per cent difference as the more significant.

3. The significance measure should be symmetric, in this sense: given a small share s and a large share S, the difference between a GDP share s and a trade share S has the same significance as the difference between a trade share s and a GDP share S.

4. For each individual region, we obtain a non-negative significance measure, zero if and only if the region's GDP and trade shares are equal. Summing over regions, we obtain a measure of difference between the two share profiles.

Finding a significance measure that has these properties is easy but not quite trivial. The absolute magnitude of the difference between the shares, for example, lacks property 2, while the ratio of the shares lacks properties 1 and 3. It is easy to see that the cross entropy has properties 1 and 2, but not 3 or 4. In particular, for an individual element i, the term $s_i^1 \log \frac{s_i^1}{s_i^0}$ may be either positive or negative.

Now if we take as our measure the symmetric cross entropy,

$$\sum_i s_i^1 \log \frac{s_i^1}{s_i^0} + \sum_i s_i^0 \log \frac{s_i^0}{s_i^1},$$

we find it has all the desired properties. In particular, the contribution of an individual element,

$$s_i^1 \log \frac{s_i^1}{s_i^0} + s_i^0 \log \frac{s_i^0}{s_i^1} = (s_i^1 - s_i^0)(\log s_i^1 - \log s_i^0),$$

is always non-negative, and zero if and only if $s_i^1 = s_i^0$. This is the notion introduced as *information discrepancy* by Kullback [9].

These ideas, as we shall see, can be applied in a wide range of contexts for a large family of data reduction tasks. They also sit well with related methods for data construction and even estimation. For lack of time and space, we shall describe here mostly easier applications of these ideas. Some of the tasks we shall describe could be performed equally well without recourse to entropy-theoretic notions; for some, entropy-theoretic ideas give help suggest the task definition; some are greatly facilitated by the entropy-theory approach. If we say of approaches to data analysis, as Larry Wall says of programming languages, that it should make easy jobs easy and hard jobs possible (Wall, Christansen and Schwartz [15]), then our emphasis here is on showing that this approach makes easy jobs easy. Towards the end we discuss briefly some ways in which it makes hard jobs possible.

We shall see that entropy concepts offer a reasonable solution to this problem. This alone does not make much of a case for employing entropy theory, since we can find other reasonable solutions with less theoretical overhead. It is however a sign as far as it goes that the theory applies in a natural way in this field; discussing it will prepare the way for entropy solutions to more complex problems.

5 GTAP services exports and imports data

We now explore the services trade data, relying heavily on the entropy-theoretic notions discussed above. We examine two arrays, one for services exports by commodity and region, and one for services imports by commodity and region. The data of a slightly later vintage than the first restricted-access prerelease of the GTAP 5 data base (we have made some bug fixes in response to problems uncovered in the course of preparing this paper), but not quite fully processed to the level where they are incorporated into the database (in the complete database, we make some adjustments to energy trade that have not been made in these data). For purposes of comparison, we shall use also analogous data from GTAP 4.

In effect, we have already reduced the five hundred odd thousand data elements for GTAP 5 services trade down to some 1560 (2 arrays by 12 commodities by 65 regions), but even this smaller number allows ample scope and motivation for data reduction.

Table 3 shows world services trade by commodity according to GTAP 5.Total world services trade is estimated at $1.2 trillion, or 19 per cent of total world trade, $6.4 trillion. The services trade total includes some $176 billion of margin services, representing international freight transport services.

We observe, for future reference:

- We did not quite tell the truth when we said that we enforced zero trade in gas distribution and water. Some trade, albeit at low levels, is recorded for those commodities. On the other hand, trade in “ownership of dwellings” is exactly zero.
- Some 77 per cent of the total is accounted for by trade and transport and “other business services”.
- We might reasonably expect to see some margin sales of insurance (representing the cost of insurance of international transactions) and trade (services associated with reexport activity). In fact we find margin sales for the transport sectors only.

Table 3: Services trade, by commodity, according to GTAP 5 ($b)

	Non-margin	Margin	Total
Electricity	12	0	12
Gas	2	0	2
Water	1	0	1
Construction	35	0	35
Trade	136	0	136
Transport n.e.c.	197	42	239
Water transport	42	109	151
Air transport	117	25	142
Communication	32	0	32
Financial services n.e.c.	37	0	37
Insurance	43	0	43
Business services n.e.c.	285	0	285
Recreational etc. services	74	0	74
Government services	54	0	54
Dwellings	0	0	0
Total	1065	176	1241

It is no surprise to find the United States the largest exporter and importer of services, since the United States is also by a considerable margin the world's largest economy. It might however be interesting to learn how regional shares in services trade differ from shares in say world GDP. More particularly, for consideration by a human reader, we would like to identify the largest or most significant or most striking differences between the trade and GDP shares.

Applying this measure to GDP and export shares, we find the most significant differences as reported in table 5. We remark that the United States and Japan earn inclusion in the table through their large absolute differences, Hong Kong and Singapore, through their large relative differences. Hong Kong shows a smaller absolute difference than Japan, but takes a higher place because of its greater relative difference; Japan shows a smaller relative difference than the Netherlands, but takes a higher place because of its greater absolute difference. The Netherlands and Belgium show similar relative differences, but the Netherlands, having the greater absolute difference, is ranked higher. Again, the Netherlands and Middle East show similar absolute differences, but the Netherlands, having the greater relative difference, is ranked as more significant. So we see that both absolute and relative difference affect the significance rankings.

Examining further table 5, two large economies, the United States and Japan, whose GDP shares exceed their services export shares, and several small open economies, Hong

Table 4: Services trade, for seven largest exporters and importers, according to GTAP 5 ($b)

Exports		Imports	
United States	207	United States	180
United Kingdom	79	Japan	117
Germany	77	Germany	111
France	76	United Kingdom	66
Japan	72	France	63
Italy	66	Italy	61
Netherlands	47	Other Near East	49
Other	614	Other	591
Total	1241	Total	1241

Table 5: Selected differences between GDP and service export shares ($\times$ 1000)

	significance	GDP share	export share
Japan	108	143	59
United States	78	275	167
Hong Kong	76	6	35
Singapore	58	3	24
Netherlands	42	12	38
Brazil	41	27	7
Belgium, Luxembourg	31	9	27
Other	*194*	*525*	*642*
Total	628	1000	999

Kong, Singapore, the Netherlands, and Belgium, whose services export shares greatly exceed their GDP shares. Finally, there is Brazil, a large country geographically, though its GDP share is not so large; we find it follows the large country rule in that its GDP share exceeds its export share. We could pursue this line of inquiry further, showing that for Japan and Brazil, the low services export share is partly due to a low share of services in total exports, while for the United States, a high share of services in total exports partly offsets a low export:GDP ratio; but we leave this for another occasion.

The relationship between GDP and services trade shares is not purely of descriptive interest, since we have at some points in the data construction process used GDP shares to proxy missing services trade shares. With the benefit of hindsight, we suspect that merchandise trade shares would prove a considerably better proxy (we considered that choice but rejected it for reasons relating to the modularity of the data construction program). One way to assess that proposition (by no means the only way!) would be to measure the information gain between the GDP shares and services trade shares, and compare it with the information gain between the merchandise trade and services trade shares. If the information gain going from merchandise trade to services trade shares is lower than the information gain going from GDP to services trade shares, then we may consider the merchandise trade shares the better proxy.

We have looked so far at the marginal totals for the services trade data: at trade by sector, or by source region, or by destination. From these data alone we could, if we had to, estimate the complete services trade data array using simple proportionality assumptions. We may call this the *pattern array*. We may ask then, how far does the actual trade data array differ from the one that we might so construct? How much information do we gain by providing the complete array rather than just the marginal totals? Where do the actual data deviate most from the pattern?

This ties in with another entropy concept, *mutual information*. One definition of mutual information is the gain in information of observing a complete second order array relative to observing only the marginal totals,

$$\sum a_{ij} \log \frac{a_{ij}}{v_i^1 v_j^2},$$

where a_{ij} are the array element shares, and v_i^1 and v_j^2 shares of marginal totals in the array total. Again, to achieve symmetry and non-negativity for individual components, we use for our significance measure the sum of the forwards and backwards information gains,

$$\begin{aligned} & \sum a_{ij} \log \frac{a_{ij}}{v_i^1 v_j^2} + \sum v_i^1 v_j^2 \log \frac{v_i^1 v_j^2}{a_{ij}} && (1) \\ = & \sum \left(a_{ij} \log \frac{a_{ij}}{v_i^1 v_j^2} + v_i^1 v_j^2 \log \frac{v_i^1 v_j^2}{a_{ij}} \right). && (2) \end{aligned}$$

We apply this measure now to the services trade data. We compare the totals for services trade by commodity and exporting region, with the data we might estimate using only the totals for services trade by commodity, and services trade by exporting region. We perform a similar exercise for services imports.

Table 6 shows for that array the most significant differences between the actual services exports data and the pattern.

The first result we remark is the unusual trade pattern for Hong Kong. Hong Kong's exports are strongly weighted toward trade and away from "other business services" and "other transport". Inspection of the full data set (not reported here) shows that Hong Kong's exports of other service commodities are far below what might be expected from the overall trade pattern. This ties in with the construction of the services trade data, which involves a unique treatment of Hong Kong reexport services. A significant part of Hong Kong's exports are actually reexports of goods from mainland China, or reexports of goods from the rest of the world to mainland China. The margins that Hong Kong earns from this activity represent a large share of its export earnings (54 per cent in preliminary GTAP 5 data). Other countries engage likewise in reexport activities, but rely on them much less for export earnings. As a special measure for Hong Kong, we estimate its reexport margins on trade to and from the Chinese mainland, and treat them as exports of trade services to the destination of the goods (the data are contributed by Shunli Yao of the University of California at Davis and Mark Gehlhar of the Economic Research Service, United States Department of Agriculture). This feature of the data set also appears on the import side, as unusually high trade services imports into China. Here the significance report has told us not something that we did not know already; it has told us something that we knew it should tell us. Overall this aspect of the trade pattern is partly artifact, since we apply a unique treatment to Hong Kong; but largely fact, since reexport services are extraordinarily important to the Hong Kong economy.

Other items in the table include an unusually high level of government services exports from the United States. Inspecting the full data set (not reported here) we find that this results from two factors; an unusually high level of exports to all destinations, and an extraordinarily high level of exports to the Middle East. Within the Middle East, most of the exports go to Saudi Arabia. We can verify that this feature exists in the source data and is not an artifact of processing. With the benefit of hindsight, and on the basis of general knowledge, we may find that this is a plausible feature of the data set.

So far we have looked at the GTAP 5 trade data set mostly in isolation. We may also compare it against the GTAP 4 data set. Since the GTAP 4 data are more highly aggregated, we need to aggregate the GTAP 5 data to the GTAP 4 classifications. Again, since there are still about one hundred thousand array elements even with the GTAP 4 classifications, we need to summarise or filter the data for human inspection.

First, we note that the value of services trade has declined slightly, from US$1.252 trillion in GTAP 4 to $1.241 trillion in GTAP 5. This may seem surprising, given growth in the world economy between 1995 and 1997, and the increasing focus on services trade. In fact, it may largely be an artifact: some $120 billion in travellers' expenditures, treated as services trade in GTAP 4, are treated as merchandise trade in GTAP 5.

Table 6: Selected differences between services trade data and pattern

		significance (× 1000)	pattern ($B)	actual ($B)
Exports				
Hong Kong	Trade	79.3	4.8	37.9
United States	Government services	19.3	9.0	25.1
United States	Water transport	18.3	25.2	9.3
Hong Kong	Other business services	18.3	10.1	1.6
Hong Kong	Other transport	15.5	8.5	1.3
Other EFTA	Water transport	13.8	2.2	9.9
United States	Trade	13.0	22.7	9.6
Philippines	Other business services	12.2	3.2	11.5
Imports				
Other Middle East	Government services	22.4	2.1	12.9
Japan	Other business services	21.9	26.9	9.2
China	Trade	17.3	4.2	15.6
China	Other business services	14.0	8.8	1.6
Canada	Other business services	7.9	8.0	17.0
United States	Construction	7.5	5.1	1.0
Italy	Other financial services	6.9	1.8	6.5
Mexico	Insurance	6.6	0.4	3.2

Next, we examine cross-entropy measures for the region shares. The greatest changes appear on the import side. In particular, we find large declines for the shares in world imports of Hong Kong (from 1.9 to 0.6 per cent) and Taiwan (from 1.7 to 0.9 per cent). This is an artifact reflecting a change in processing between GTAP 4 and GTAP 5. In both data bases, we have no services trade data for Hong Kong or Taiwan (nor indeed for many other countries). For GTAP 4, where data are lacking, we fill them in using shares in world merchandise trade; for GTAP 5, we use shares in world GDP. It so happens that for Hong Kong and Taiwan, the merchandise trade shares far exceed the GDP shares. Between now and the final release, we will reassess the new policy, and perhaps revert to the old. Clearly however the only satisfactory solution is to obtain actual statistics.

The next most significant change is for imports into China, and here the import share rises between GTAP 4 and 5 (from 1.9 to 3.1 per cent). Here again the explanation lies in the different processing of the two tables. In GTAP 4, a large share of China's total services imports was initially assigned to the commodity electricity; later, other information was imposed setting China's electricity imports to a low value; the net effect was to remove a large part of China's services imports. In GTAP 5 we never allocate services trade to electricity, so we preserve the initial import total more closely through processing. A similar explanation applies to the increase in Denmark's share in world imports, from 0.5 to 1.0 per cent.

For EFTA, we see a decline in the export share from 3.6 per cent in GTAP 4 to 2.6 per cent in GTAP 5. This again is artifact rather than fact. For GTAP 4 our source data provided a services export total for Switzerland, the largest EFTA member economy, but for GTAP 5 they do not. Consequently, in GTAP 5 we filled in Switzerland's data using GDP shares. Comparison with other information would suggest that this led to severe underestimation of Switzerland's services exports, and consequently to EFTA's. Again, this highlights an area for action before the final release.

For the sector shares, an entropy-based report is really redundant, since they are few enough that we can readily inspect them all. Still, it is interesting to see what the entropy method yields. Table 8 shows the results.

As we see, the main result from the comparison is the fall in the share of government services, from 10.6 to 4.3 per cent. This may reflect an error in data construction in GTAP 5; as we now understand, we have excluded from government services some trade which rightly belongs there, education and health services consumed abroad. The other major feature, a rise in the share of "business etc. services", from 26.3 to 35.4 per cent, appears largely to reflect genuine differences in the source data, between the input-output tables from which the sector splits were taken in GTAP 4, and the IMF balance of payments statistics from which they are taken in GTAP 5; though discrepancies in sectoral definitions may also play a minor role (causing some slight overestimation of the share in GTAP 5).

We now proceed further, to compare between GTAP 4 and GTAP 5 not just the first-order trade subtotals, trade by source region or destination region or commodity, but second-order subtotals such as trade by source region and commodity. We want to do this in a way

Table 7: Major differences in region shares between GTAP 4 and GTAP 5 ($\times$ 1000)

	significance	GTAP 4 share	GTAP 5 share
	Exports		
EFTA	5.0	36	26
Other Middle East	2.6	17	23
Malaysia	2.6	7	11
Philippines	2.2	8	11
Former Soviet Union	2.2	16	21
Viet Nam	1.9	2	1
India	1.8	5	7
	Imports		
Hong Kong	19.1	19	6
China	8.3	19	31
Taiwan	7.3	17	9
Denmark	4.7	5	10
Germany	4.4	107	90
Philippines	3.6	6	11
United States	2.1	131	145

Table 8: Major differences in sector shares between GTAP 4 and GTAP 5 ($\times$ 1000)

	significance	GTAP 4 share	GTAP 5 share
Government services	81.6	106	43
Business etc. services	38.6	263	354
Construction	6.4	18	28

that does not repeat the results of the first-order subtotal comparisons. For example, having found that the ”business etc. services” share rises globally between GTAP 4 and GTAP 5, we do not want to report this fact again for all major exporting regions. Instead we want to report those regions where exports of ”business etc. services” have risen more than would have been expected from the global increase in the business services share. To use a terminology introduced earlier, we want to focus now on changes in mutual information between regions and sectors.

We proceed as follows. We construct an estimated GTAP 5 exports array (with dimensions, exporting region by commodity), taking as our starting point the GTAP 4 exports array. We adjust the GTAP 4 array so that the marginal totals, exports by region, and exports by commodity, agree with the corresponding totals for GTAP 5. We then calculate our entropy-based significance measure for discrepancies between the estimated and actual GTAP 5 arrays. This gives us a report on differences between the GTAP 4 and GTAP 5 arrays that cannot be attributed to changes in the overall region and sector shares. We proceed similarly for imports; the results are shown in table 9.

From table 9, we see that for several countries, exports of construction services have risen from negligible to substantive levels between GTAP 4 and 5. This applies to Japan, Turkey, and Sweden, and to some other regions not reported in the table. This reflects differences in the source data; for GTAP 4, the source input-output tables for those regions showed negligible or zero exports of construction services, while for GTAP 5, the IMF balance of payments data shows substantive export values. This may represent differences in concepts and classifications between the data sources, or inaccuracies in one or both sets of sources. We may rule out a third possible explanation, genuine changes through time, since the IMF balance of payments data show no radical change in construction services exports between the two data reference years 1995 and 1997.

Returning to a theme from earlier tables, we see increases in exports of government services from the United States (from 0.6 to 2.0 per cent of world services trade), and imports of government services into the Middle East (from 0.1 to 1.0 per cent). This is plausibly attributable to improvements in source data. We may note here that for GTAP 4, we had to adjust drastically the data for government services trade for the United States—not because the trade levels reported in the US input-output table were too small, but because they were so enormous that even if we allowed the US to monopolize the trade, its government services exports exceeded all possible trading partners’ total government services imports; and likewise for US imports and partner country exports. For the Middle East outside Morocco, we had in GTAP 4 no good statistical basis for its services trade.

Likewise the increase in the European Union’s exports of trade and transport services (from 8.9 to 13.9 per cent), and the fall in its exports of “business etc. services” (from 13.2 to 8.8 per cent) reflect source data changes. Since both the input-output tables (GTAP 4) and the IMF balance of payments data (GTAP 5) appear reasonable data sources, it is not clear which representation is preferable.

The decline in imports of construction services into Indonesia appears a plausible can-

Table 9: Selected differences between GTAP 4 and GTAP 5 in (region, commodity) shares ($\times 1000$)

		significance	estimated	actual
		Exports		
Japan	Construction	60.2	0.0	5.4
Other European Union	Trade and transport	33.2	88.8	139.6
Other European Union	Business etc. services	25.7	131.9	88.0
United States	Government services	23.1	6.4	20.2
Turkey	Construction	17.9	0.0	1.6
Sweden	Construction	17.1	0.0	1.5
Singapore	Business etc. services	14.4	3.8	12.4
Other North Africa	Business etc. services	14.1	0.2	3.6
Central Eur. Assocs	Construction	13.4	6.1	1.0
Other North Africa	Government services	13.0	4.3	0.4
		Imports		
Indonesia	Construction	55.8	6.7	0.0
Japan	Construction	36.5	0.1	5.5
Other Middle East	Government services	31.1	1.0	10.4
Sweden	Construction	21.7	0.0	2.3
Other Middle East	Trade and transport	16.0	26.6	12.2
United States	Business etc. services	14.8	64.2	41.2
Germany	Government services	14.6	7.3	1.4
Other sub-Saharan	Business etc. services	13.1	0.3	3.9
Other Central Am'n	Construction	12.7	3.2	0.2
Indonesia	Business etc. services	12.7	1.0	5.9

didate for genuine change over time, given the economic crisis in Indonesia and some other East Asian economies in 1997. Sadly, it appears to reflect a deficiency in the IMF source data; Indonesian construction services imports are not reported for any year, whether before or during the crisis.

6 Making hard jobs possible

We have shown a several applications of entropy theoretic methods, mostly to tasks of data reduction. These have mostly been relatively simple tasks, verifying that this approach does at least make easy jobs easy. We now touch briefly on its potential for making hard jobs possible.

Here are some harder tasks that entropy-theoretic methods can address.

- Given a third-order array, say services trade exports by year, commodity, and country, extract from it a pattern combining the following factors:

  - a country-specific factor (country shares),
  - a commodity-specific factor,
  - a year-specific factor,
  - a (country,commodity)-specific factor,
  - a (country,year)-specific factor, and
  - a (commodity,year)-specific factor,

  such that the (country,commodity)-specific factor does not affect the overall country or commodity shares, and likewise for the other second-order factors. This exercise yields a richer description of the data than the first-order shares alone. For example, the (country,year) factor lets us identify countries that have grown faster or slower than the average, and does so in a way that abstracts from from any compositional effects that arise if the country has for example a trade pattern weighted towards a commodity that grows at a rate different from the average. We can then identify elements of the original array that stand out against this rich pattern: for example, a data value that is unexpectedly high even after allowing for the possibilities that the country's trade pattern is biased toward that commodity, or that the country's trade level is especially high in that year, or that trade in that commodity is especially high in that year.

- Extract from a third-order array a rich pattern as described above, but do it even though some data elements are missing, and without allowing the missing elements to bias the pattern (for example, if there are many missing data elements for slower-growing countries in later years, do not allow that to bias the pattern). Use the resulting pattern to fill in the missing values.

Table 10: Data sets used in constructing GTAP non-margin trade data

description	(1)	(2)	(3)
merchandise trade	N	Y	N
reexport services	N	N	Y
non-margin services	Y	N	Y
travellers' expenditures	Y	Y	Y

(1) from services trade data set?
(2) affects non-services trade data in GTAP?
(3) affects services trade data in GTAP?

- Do as above, but this time instead of dealing with missing values, deal with data that have been aggregated in a non-systematic fashion, so that for example some countries provide fully disaggregated data and some provide aggregated data, and different aggregate commodity classifications are used for different countries and different years. Use the pattern to disaggregate the data set.

7 From the services trade data set to services trade data in GTAP

We offer now a few belated words about the services trade data set itself, and its relation to services trade data in the GTAP data base.

That relation is more complex than one might expect: not all services trade data in the GTAP data base comes from the services trade data set, and not all GTAP data that comes from the services trade data set is GTAP data (table 10).

Specifically, the reexport services data set contains estimates for services used in Hong Kong to facilitate flows of goods to and from mainland China. These services are represented as exports of trade services from Hong Kong to the destination of the goods (China for goods imported into China, other countries for goods exported from China). The reexport services data set is supplied by Shunli Yao and Mark Gehlhar (Yao [16]), and is independent of the services trade data set.

Before any of the services trade data are brought into the GTAP data base, they must be aggregated from countries to GTAP regions. All components of the services trade data set are defined over a standard set of 211 countries. Typically about data for about half of those countries is not available in the source statistics, but is constructed synthetically. This is not as serious as it may sound, since most missing data are for small countries; so that typically again, more than 90 per cent of the total value in the 211-country array derives from source statistics.

To some extent we use merchandise trade data in constructing the services trade data. In particular, there is a balance requirement for the trade data, that the total value of margin services should be equal to the total CIF value of non-margin trade, less the total FOB value. We calculate the CIF-FOB difference from the merchandise trade data, then balance the margin services usage and supply arrays against it.

In constructing the margin services usage array we use both margin imports data from the main services trade data set, and a margin usage data set supplied by Mark Gehlhar. The margin imports data does not indicate the transported commodities to which the margins apply, nor the transported commodities' country of origin. Mark Gehlhar's data set does contain that information, but unfortunately applies to United States trade only. To construct the initial margin services usage matrix we heroically extrapolate the modal shares in the Gehlhar data from US trade routes to all trade routes. In so doing we adjust shares to take account of whether trade partners share a common border, and whether they are on the same continent (so that trade between countries on different continents tends to have low land transport shares). We then balance the margin usage array so that the total margin on each merchandise trade flow agrees with the merchandise trade CIF-FOB value difference, but the composition of margin usage on total imports into each country agrees with the margin imports data from the main services trade data set.

One major category in the services trade source statistics is travel, also described as travellers' expenditures. This corresponds to the GATS supply mode *consumption abroad*. In a CGE framework, this may be considered not as a commodity but as a final demand category (on the export side) or a source of supply (on the import side). Unfortunately the source statistics currently at our disposal do not indicate how travellers' expenditures should be broken down across ordinary goods and services. The split used in the current data set is wholly conjectural. We know already that the split is inaccurate in that it allows a zero share for "government services", whereas in fact education services supplied to foreign students and health services supplied to foreign patients are recorded here. Statistics on education and health services consumed abroad may be available in IMF balance of payments supplementary information; we have not as yet acquired that information.

8 Work ongoing on services trade data

Possible directions for work on services trade data in the short- to medium-run include:

- using merchandise trade rather than GDP shares to proxy missing services trade shares;
- filling in missing data for the reference year, 1997, using data for 1995 and 1996;
- exploiting the EUROSTAT bilateral trade data;

- seeking data for the allocation of travellers' expenditures across commodities,
- documenting and publishing the services trade data set underlying the services trade data in the GTAP data base.

References

[1] M. Bacharach. *Biproportional Matrices & Input-Output Change*. Number 16 in University of Cambridge Department of Applied Economics Monographs. Cambridge University Press, 1970.

[2] J.P. Burg. *Maximum Entropy Spectral Analysis*. PhD thesis, Department of Geophysics, Stanford University, 1975.

[3] Mark Gehlhar. *Reconciling merchandise trade*, chapter 11.A. In McDougall et al. [11], 1998.

[4] A. Golan, G. Judge, and D. Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Wiley, Chichester, 1996.

[5] A. Golan, G. Judge, and S. Robinson. Recovering information in the case of underdetermined problems and incomplete data. *Review of Economics and Statistics*, 76:541–9, 1994.

[6] E.T. Jaynes. Informational theory and statistical mechanics. *Physics Review*, 106:620–30, 1957.

[7] J. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35:917–26, 1956.

[8] A.N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions in Information Theory*, IT-14:662–4, 1968.

[9] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[10] Mukela F. Luanga and Wusheng Yu. Trade in services data: Data source, classification, evaluation, and construction. Technical report, Center for Global Trade Analysis, Purdue University, 1999.

[11] Robert A. McDougall, Aziz Elbehri, and Truong P. Truong, editors. *Global Trade, Assistance, and Production: The GTAP 4 Data Base*. Purdue University, 1998.

[12] M.H. Schneider and S.A. Zenios. A comparative study of algorithms for matrix balancing. *Operations Research*, 38:439–55, 1990.

[13] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–659, 1948.

[14] H. Theil. *Economics and Information Theory*. North-Holland, Amsterdam, 1967.

[15] Larry Wall, Tom Christiansen, and Randal L. Schwartz. *Programming Perl*. O'Reilly, Sebastopol, CA, 1991.

[16] Shunli Yao. *Estimating Hong-Kong Re-export Markups*, chapter 11.B. In McDougall et al. [11], 1998.