

## **Report No. 3 (R3)**

# **Generation of a U.S. Commodity Flows Matrix Using Log-Linear Modeling and Iterative Proportional Fitting (Freight Analysis Framework)**

### **Synopsis**

This paper describes the task of constructing a set of commodity class and mode specific annual origin-to-destination flows for the entire United States, covering to the extent feasible all domestic as well as all imported and exported goods. The task takes as its starting point the 2002 U.S. Commodity Flow Survey and a number of supplemental data sources out of which a single commodity flow matrix is constructed. The product of this effort is a four dimensional matrix of flows that can be reported in annual tons, annual dollar value, and annual ton-miles, with the principal dimensions being:

- shipment origination region (O)
- shipment destination region (D)
- the class of commodity being transported (C), and
- the mode of transportation used (M)

Having identified the problem to be solved, the paper describes how a log-linear modeling approach can be used to estimate missing values in the initial flow matrix, and how subsequent application of iterative proportional fitting can be applied to both reported and model-estimated flows to generate a complete O-D-C-M matrix that meets reported marginal flow totals. Preliminary results are presented for portions of the full flow matrix. Appealing features of the method include its ability to make full use of the available freight movement data, and its ability to draw in data from not other mode and commodity specific data sources where deemed useful to the estimation/data gap filling process. The paper also describes how the approach deals with the different types of zero valued cells found in large, sparse matrices, and proposes an approach to model validation in the absence of comparative data from other sources, especially where truck shipments are concerned. An appendix provides a fully worked numerical example of the log-linear modeling methodology used in estimating missing cell values.

## 1. Purpose

This paper describes and presents a numerical example of the commodity flow matrix gap-filling technique being applied to U.S. commodity flow data within the Freight Analysis Framework (FAF2) project. The Office of Operations, Freight Management and Operations, within the Federal Highway Administration (FHWA, U.S. Department of Transportation (USDOT) funded the project. The technique offers a flexible, reproducible and statistically based method for estimating missing cell values (i.e. missing origin-to-destination flows) in large interaction matrices containing many empty cells. The principal data product of the project is a set of estimated annual flows of goods between all pairs of FAF geographic regions, by 43 specific classes of commodity and by specific transportation modes and mode combinations. Flows are reported in terms of annual tonnages transported, their associated dollar values, and ton-miles of activity. Commodity classes, transportation modes and geographic regions are all based on breakdowns used within the 2002 U.S. Commodity Flow Survey (CFS).<sup>1</sup>

The need for data modeling described in this paper comes from the limited coverage that the CFS is able to provide, in terms of both the scope of the industry/commodity sectors it captures and also because the survey's limited sample size prevents robust estimation of many of the origin-destination-commodity-and-mode (O-D-C-M) specific flow totals needed for freight planning studies. The data modeling approach described in the paper seeks to compensate for these weaknesses in the U.S. multimodal commodity flow picture by making full use of the statistical data available from:

- the CFS and
- other commodity flow datasets offering national, commodity specific, and/or mode specific coverage.

Non-CFS data are used in two different ways, as shown in Figure 1. The most important use of these data is as a supplement to the CFS movement estimates within each of its 43 commodity classes. A great deal of additional commodity movement data is needed to complete the annual U.S. commodity flow picture. This “out-of-scope” (to the CFS) component of the FAF flows matrix constitutes over 40% of all freight moved within the nation. This includes freight moved by the following industrial sectors (all of which fall under one of the original 43 FAF/CFS commodity classes):

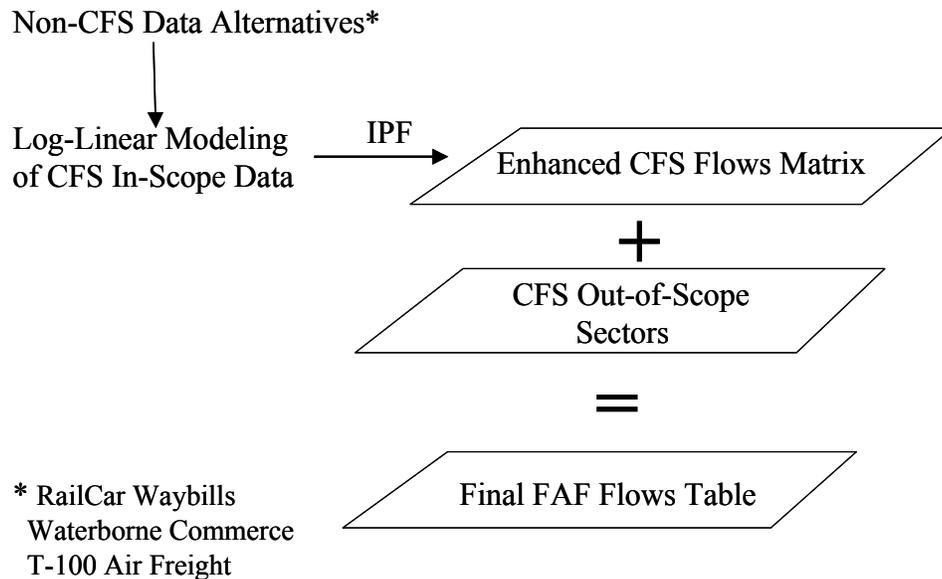
- farm-based shipments
- crude petroleum
- in-transit goods
- municipal solid waste products
- logging

---

1. Census Bureau (2004) 2002 Commodity Flow Survey. Report EC02TCF-US. US Department of Commerce, Economics and Statistics Administration, Suitland, MD.  
<http://www.census.gov/econ/www/cfsnew.html>

- fisheries
- retail
- construction
- services
- publishing
- government
- household and business moves

Also used extensively, both as alternative sources of U.S. commodity flows and as a means of validating the FAF flows matrix, are a number of mode specific sources of based on nationwide carrier surveys. These data are referred to in Figure 1 as “non-CFS data alternatives” since they are used in the current approach to fill gaps in CFS coverage where it is clear that a sampling zero existed in the survey. For example, significant rail movement in the railcar waybill sample where a CFS zero valued cell exists is taken to indicate a CFS sampling rather than a CFS structural zero. These alternative data sources are in all cases data that cover many of the same flows captured by the CFS but reported in different ways and with different holes in them for FAF matrix construction purposes.<sup>2</sup>



**Figure 1. CFS In-Scope and Out-of-Scope Components of the FAF Flows Matrix**

Alternative data sources include the U.S. Surface Transportation Board (STB) annual railcar waybill dataset<sup>3</sup>, the U.S. Army Corps of Engineers (USACE) waterborne

<sup>2</sup> In particular, being mode-dedicated, they do not provide true shipment origination and destination points: not containing any evidence of truck draying, often over significant distances and across regional boundaries, to airports, seaports, or other intermodal terminals

<sup>3</sup> [http://www.stb.dot.gov/stb/industry/econ\\_waybill.html](http://www.stb.dot.gov/stb/industry/econ_waybill.html)

commerce dataset<sup>4</sup>, and the Bureau of Transportation Statistics' T-100 Domestic and International Air Freight dataset. Each of these datasets has its weaknesses, but all provide a level of coverage whose scope is all commodities moved within each mode. The waterborne commerce data in particular are premised on a 100% sample of carrier responses, as are the air-freight data, although this dataset lacks commodity detail and reports only total revenue tons transported. The railcar waybill sample does an excellent job of capturing the largest rail movements in considerable commodity detail. Also used in this matrix enhancement process was a U.S. Census Bureau matrix of reported sample counts for each O-D-C-M cell in the CFS matrix. These data offered some additional insight into the presence of empty or very low valued cells that is difficult to extract from the publicly available CFS data tables.

The methodology described in this paper allows these and other mode and commodity specific datasets to be used to enhance further the U.S. commodity flow picture. Commodity specific datasets such as the Energy Information Administration's (EIA) data<sup>5</sup> on annual U.S. coal shipments are good candidates for this exercise. Each of these data sources is described in detail in supporting FAF<sup>2</sup> project documentation. They are also critiqued in references [1] through [4].

Especially challenging is the problem posed by U.S. import and export shipments. The former are out of scope for the CFS, but it is unclear where such shipments cease to be imports and become U.S. company-based movements, and are therefore already being captured, at least in part, within the CFS domestic movements sample frame. The unanswered question within currently available data sources is whether these commodity movements change ownership at or near the U.S. seaport of entry, or after a lengthy inland transit. While export shipments by U.S. companies are a part of the CFS sample frame, known discrepancies between these data and other sources of U.S. foreign trade data raise questions about CFS robustness as a region-to-region data source. This means that modeling of both U.S. imports and exports (as well as in-transit shipments that cross U.S. territories but are never technically U.S. owned commodities) poses a non-trivial task when trying to fill in the complete U.S. freight movement picture. Imports and exports are discussed in more detail in Section 5.

First, Section 2 defines the working dimensions of the FAF2 commodity flow matrix. Section 3 then describes the two principal methods used to construct the FAF flow matrix from the above referenced data sources: the log-linear modeling and iterative proportional fitting (IPF) routines shown in Figure 1. Section 4 provides the initial results of applying these routines to the above datasets. Sections 6 and 7 of the paper then discuss, respectively, model validation issues and ongoing extensions to the current flow matrix construction effort

---

<sup>4</sup> <http://www.iwr.usace.army.mil/ndc/wcsc/wcsc.htm>

<sup>5</sup> <http://www.eia.doe.gov/cneaf/coal/page/coaldistrib/coaldistrib.html>

## **2. Problem Statement**

### **2.1 Problem Dimensions**

The complete FAF 2002 U.S. Commodity Flows Matrix contains 138 x 138 origin-to-destination (O-D) region shipments, broken down by 43 commodity classes and by 7 major mode/mode combinations.

Table 1 lists the 43 FAF commodity classes and the 7 modes and mode combinations from which the initial FAF flows matrix has been created. Both sets mirror the classes used in the 2002 CFS. In total this represents the creation of a four-dimensional intra-U.S. commodity flows matrix containing 5,732,244 cells (138 x 138 x 43 x 7). While the majority of these cells contain zero valued flows for “structural” reasons (e.g. no

**Table 1. FAF Commodity & Mode Disaggregations**

SCTG

<b>Code</b>	<b>Commodity Classes</b>	<b>Transportation Modes</b>
01	Live animals and live fish	01 Truck
02	Cereal grains	Private truck
03	Other agricultural products	For-hire truck
04	Animal feed and products of animal origin, n.e.c.	02 Rail
05	Meat, fish, seafood, and their preparations	03 Water
06	Milled grain products and preparations, and bakery products	Shallow draft
07	Other prepared foodstuffs and fats and oils	Great Lakes
08	Alcoholic beverages	Deep draft
09	Tobacco products	04 Air (inc. truck-air)
10	Monumental or building stone	05 Truck-Rail Intermodal
11	Natural sands	06 Other Multiple Modes
12	Gravel and crushed stone	Including:
13	Nonmetallic minerals n.e.c.	Parcel, USPS or courier
14	Metallic ores and concentrates	Truck-water
15	Coal	Water-rail
16	Crude Petroleum	07 Other and Unknown modes
17	Gasoline and aviation turbine fuel	Including pipeline
18	Fuel oils	
19	Coal and petroleum products, n.e.c.	
20	Basic chemicals	
21	Pharmaceutical products	
22	Fertilizers	
23	Chemical products and preparations, n.e.c.	
24	Plastics and rubber	
25	Logs and other wood in the rough	
26	Wood products	
27	Pulp, newsprint, paper, and paperboard	
28	Paper or paperboard articles	
29	Printed products	
30	Textiles, leather, and articles of textiles or leather	
31	Nonmetallic mineral products	
32	Base metal in primary or semifinished forms and in finished basic shapes	
33	Articles of base metal	
34	Machinery	
35	Electronic and other electrical equipment and components and office equipment	
36	Motorized and other vehicles (including Parts)	
37	Transportation equipment, n.e.c.	
38	Precision instruments and apparatus	
39	Furniture, mattresses and mattress supports, lamps, lighting fittings, etc.	
40	Miscellaneous manufactured products	
41	Waste and scrap	
43	Mixed freight	
—	Commodity unknown	

Key: n.e.c.=not elsewhere classified.

flows of commodities out of regions that do not produce them, no truck or rail trips into and out of Hawaii, etc.), there are still a significant number of cell values that need to be estimated: and even a matrix requiring 5% of the cells to be filled requires an estimated value for over half a million cells. Not all missing cell values need to be solved for at once: although a single approach that captures all statistical effects is certainly preferable.

To better model truck shipments, the mode with the largest data gaps, the procedure incorporates an additional, shipment distance-based dimension to help with the cell value estimation process (see below). An additional, temporal dimension that combines data from the 1993, 1997 and 2002 CFS surveys may also prove valuable in the future for generating more robust cell values across all mode/commodity categories. The initial FAF matrix uses 2002 dated CFS data only.

## 2.2 Nature of the Missing Data

**CFS Data Tables** A combination of data suppression for confidentiality reasons, limited sample size, and limitations to the scope of the CFS (across industrial sectors) mean that many cells that ought to contain a flow are empty. The questions that require answers are what size each of these flows should be and which cells ought to contain a positive flow at all. It should also be noted that even when the data are summed across a particular row or column in the CFS O-D-C-M matrix, sometimes data are missing data for these two as well as three dimensional margins, and not just data on the four-dimensional flows sought. A study of the complete set of 2002 CFS data products indicates that there are a good many data matrices to work with. This includes the most detailed of the published matrices, Table 17, which reports annual tons, dollar value and ton-miles shipped by state of origin, state of destination, mode and 2-digit (43) commodity classes.<sup>6</sup> Other tables provide 1, 2 and 3 dimensional looks at this same data, including flows broken down to the 114 CFS/FAF intra-US geographic regions of interest. Without going through the contents of each data table the gaps in current CFS coverage can be summarized as follows:

- commodity specific annual shipment generation and attraction totals exist but there are no origin-to-destination (O-D) flow estimates, either by mode or summed over all modes
- total annual O-D commodity flow estimates exist but without any modal breakdown.
- modal share estimates exist but lack the geographic and/or commodity detail required of the FAF flows matrix.
- data on shipment lengths exists, by mode and/or commodity, but with little or no linkage to O-D geography.

---

<sup>6</sup> <http://www.census.gov/svsd/www/02CFSdata.html>

That is, a flow matrix is available that contains a variety of levels of coverage of its 1, 2, 3 and 4-dimensional data elements, with many gaps in it.

### 3. Matrix Generation Methodology

#### 3.1 Requirements of the Method

An ideal method for filling missing cells in the FAF flow matrix is considered to display the following characteristics:

1. the ability to make the most use of existing data within the matrix in the estimation of missing cell values
2. the ability to bring different, including non-CFS sources of flow estimates into the solution, including completely new one, two and three-dimensional data tables, as needed
3. the ability to fill in missing cell values while maintaining reported marginal flow totals and observed cell values across all dimensions of the matrix
4. the ability to handle missing values at multiple levels of data aggregation

Although a number of gap-filling methods exist, a combination of iterative proportional fitting and log-linear modeling (including spatial interaction modeling) was found to offer each of the above features. This approach can also be used to update the flow matrix on an annual or longer-range basis given reported or forecast changes in marginal flow totals. The method used expands the FAF matrix problem by three additional dimensions, as described below.

#### 3.2 Iterative Proportional Fitting

Automated methods for estimating cell values in large and multi-dimensional matrices often involve some form of iterative proportional fitting, or IPF: a matrix-balancing technique that has been in use for over half a century [5, 6]. Consider the simplest two-dimensional case, in which  $O(i)$  and  $D(j)$  are a set of row ( $i$ ) and column ( $j$ ) totals respectively (e.g. annual freight tons produced at each  $i$  and consumed at each  $j$ ), and where  $T(i,j)$  = the tons of freight shipped from region  $i$  to region  $j$  annually.

Mathematically, a simple IPF routine applied to this problem can be stated as:

$$T(i, j, r+1) = T(i, j, r) / \sum_j T(i, j, r) * O(i) \quad (1)$$

$$T(i, j, r+2) = T(i, j, r+1) / \sum_i T(i, j, r+1) * D(j) \quad (2)$$

where  $r$ ,  $r+1$  and  $r+2$  refer to successive iterations, and where equations (1) and (2) can be applied iteratively until at some iteration  $r+g$  is gotten:

$$\sum_j T(i, j, r+g) = O(i) \text{ for all } i, \text{ and } \sum_i T(i, j, r+g) = D(j) \text{ for all } j \quad (3)$$

such that the  $T(i,j)$  cell estimates fit with all of the flow marginal totals.

IPF is often used in such cases when reliable cell estimates cannot be obtained directly, but estimates of the variables of interest are available at a higher level of aggregation. This is exactly the case described above for the FAF O-D-C-M matrix. The idea behind IPF is to seed each of the missing data cells with an initial estimate of some form, then iterate over all of the different margins of the matrix until a new balance has been obtained that does the least damage to the estimates in the rest of the matrix, and while retaining the values of the statistically more robust (typically) marginal totals that often represent the reported data. The approach is especially appealing as an application to commodity flow modeling. It is possible to take advantage of common traits, such as distance decay and the preference for using certain modes to handle certain shipment distances and commodity types, to develop intelligent missing element models. Of particular interest for FAF purposes is the combination of IPF with hybrid forms of log-linear model, including spatial interaction models based on the derivation of “maximum likelihood” estimates of the missing cell values.

### 3.3 Log-Linear Modeling of Missing Flows: Adding an “Alternative Data Model” Dimension

Numerous practical examples of applying IPF-based methods exist, including some directly relevant applications to multi-dimensional movement tables [7, 8, 9, 10]. An example model may help. In this example a single commodity class is assumed for brevity and to solve for the other three (O-D-M) dimensions. When a region generates freight traffic, it is referred to as region “i;” and when a region receives this traffic, it is referred to as region “j.” Individual modes are designated “m.” The data product sought is a fully filled matrix of freight flows, measured in annual tons moved,  $\{F_{ijm}\}$ , broken down across each of these three dimensions. For this given commodity an estimate is made of the following “fully saturated” multiplicative model of the tons shipped from region i to region j by mode m as:

$$F_{ijm} = \tau_0 \times \tau_i^O \times \tau_j^D \times \tau_m^M \times \tau_{im}^{OM} \times \tau_{jm}^{DM} \times \tau_{ij}^{OD} \times \tau_{ijm}^{ODM} \quad (4)$$

which we would solve computationally using logs as:

$$\ln F_{ijm} = \theta + \lambda_i^O + \lambda_j^D + \lambda_m^M + \lambda_{im}^{OM} + \lambda_{jm}^{DM} + \lambda_{ij}^{OD} + \lambda_{ijm}^{ODM} \quad (5)$$

The various  $\lambda$ 's, often termed the model effects, are a set of model-estimated parameters that will return the original cell estimates. (The numerical example given in the Appendix shows how to compute these parameters). For example, the  $\lambda_{ij}^{OD}$  effect returns the impacts of O-D separation on the resulting cell estimate, while  $\lambda_i^O$  represents the size of origin effect. Given a completely filled in flows matrix, equation (5) will reproduce the

cell estimates exactly. There is interest, for FAF purposes, in how such a model performs with missing data.

***Handling Zero-Valued Cells:*** The FAF flow matrix is a very large and very sparse multi-dimensional table. That is, it contains a large number of zero valued cells. Determining which of these cells are true or structural zeros, and which are zero valued is needed because of the limitations of CFS sampling, termed sampling zeros. In the CFS there are also many cells containing either an “S” or a “D” flag in the published tables: the S signifying a cell estimate too poor in quality to be reported (including all cells with a coefficient of variation of greater than 50%), the D signifying a need to suppress the cell value to avoid disclosing data about individual company activity.<sup>7</sup> While such suppression causes the loss of data it also provides information that can be used in generating the cell-specific effects sought. (It also provides an additional rationale for carrying out this type of log-linear modeling). Where an “S” or a “D” occurs, there exists a positive flow value for that cell. Therefore, use of the log-linear modeling technique, described above, is needed to generate a suitable effect, and subsequently a positive flow estimate, for such a cell.

But how good is the estimate of the size of these “empty” cells? In some cases this can be quite a large value, because a coefficient of variation (CV) of over 50% does not necessarily mean that only small O-D flows need to be dealt with. For example, it may imply a small sample containing a one very large flow and a number of smaller flows of that particular mode/commodity combination. To improve the ability to identify and estimate these missing cell values two additional data modeling steps have been taken:

(1) First, information was obtained from the CFS Branch of the U.S. Census Bureau that identifies whether any sample responses, of whatever quality, were received for each cell in the O-D-C-M flow matrix, as part of the 2002 CFS. Where zero responses were obtained, these cells were initially treated as though they are structurally zero for the purposes of log-linear model effects generation and subsequent cell estimation via IPF. In some cases this may mean losing small volume O-D flows that might have been caught by another survey, but the assumption is that the size of such a flow is going to be quite small and for our purposes “under the radar”.

(2) To further verify this assumption, as well as bring more qualitative information into the process, a search was done on possible reporting of flows within a specific O-D-C-M cell, among other databases. In particular, 2002 STB railcar waybill data were examined as well as USACE 2002 waterborne commerce O-D-C data, BTS 2002 air freight data, and other regional, commodity or industry specific sources, such as the Energy Information Administration’s (EIA) 2002 data on annual coal shipments (each dataset suitably modified to match FAF regions and commodity classes). It was determined whether a specific O-D-C-M cell in the FAF flow matrix has a flow estimate reported in *any* of these databases; and if so, then it is treated as a value to be filled in by the log-linear model-based cell estimation procedure. This procedure is now described.

---

<sup>7</sup> There are no “D” cells reported in the 2002 CFS, only “S” cells. Both types of suppression occur in the 1993 and 1997 surveys.

The log-linear modeling process begins by assigning empty cells a value of = 1.0 (log = 0.0). Then, the additive equation (5) is applied from which we obtain an estimate of each cell's missing value based only on the remaining, reported cell values. This estimate is further refined by introducing additional data into the problem, such as data from the railcar waybills. Here this waybills data acts as a second estimate, or "alternative data model," of the rail flows in each commodity class (cf. Figure 1). At this point, consideration is given to combining these data in a number of ways. The decision was made to treat the waybill flows as though they were a separate dimension, or a second set of commodity-specific tables, reproducing the rail portion of the CFS-based FAF flows matrix. This allowed filling in the missing valued cells using a combined CFS and waybills-inclusive log-linear model. The best way to combine these data is not obvious, however. It includes the use of simple additive weighting schemes such as  $[(1-w)*CFS + w*Non-CFS]$ , or more selective weighting schemes applied only to cells with questionable CFS values. The project's initial use of the waybill data was of this latter type, pending a more defensible form of general weighting process. In this approach those CFS cells with zero values, but for which the waybills report a flow taking place, are assigned a positive effects value by the log-linear model. The subsequent application of the full IPF routine to these seeded effect cells than fills in a maximum likelihood value for that cell, subject to the appropriate CFS marginal control totals.

A similar operation is also being carried out on those parts of the FAF O-D-C-M matrix involved with water and air freight transportation, using Corps of Engineers (USACE) and BTS/Office of Airline Information (BTS/OAI) data respectively as the second "model" of these mode-specific flows. In practice this means putting each of these data sources into its mode specific slice of what is termed the "alternative data model" dimension of the expanded FAF matrix. By housing these alternative modal data sources within a single dimension of the matrix in this manner, this allows, without loss of generality, for the application of more sophisticated across the board CFS + non-CFS weighting schemes in the future.

It would also be preferable to carry out a similar operation for truck shipments, but in this case a lack of suitable data means that we need an alternative solution. The approach being used is described below.

### **3.4 Incorporating Shipment Distance Information: Adding a Sixth Dimension**

Two alternatives were considered for enhancing the CFS truck movement matrices: (1) substitute a spatial interaction (SIA) model of flows in place of a second "data model;" or (2) add a sixth dimension to the FAF flow matrix using commodity specific shipment distance interval data from the CFS to force truck shipment volumes to comply with these implied ton-mileage totals. Given the level of effort and potential difficulty involved in fitting a series of spatial interaction models to the FAF truck flows this second approach was adopted for initial matrix construction. This introduced a sixth dimension into the log-linear modeling solution, based on CFS reported tons within mileage ranges of less than 50, 50-99, 100-249, 250-499, 500-749, 750-999, 1000-1499, 1500-2000, and over

2000 miles.<sup>8</sup> Using this distance-interval data, distance-decay information is brought directly into the flow matrix adjustment process and avoids the need to assume a specific parameter value (or generate a set of parameter values) to fit each commodity specific SIA model. While the approach was thought to be necessary in order to control the resulting allocation of otherwise unreported truck flows, it was valuable to apply the method across all modes in the same manner, adding further “known” constraints to the estimated FAF flow matrix.

Note also that this approach does not preclude future use of commodity specific interaction models to fill in the truck slice of the above described alternative data model dimension: as long as the constraints on distances traveled within the full FAF matrix are made to comply with distance interval-based tons, dollars, and reported ton-mileage totals, an operation taken care of by the IPF routine for any given set of marginal totals. Indeed, commodity specific modeling of such flows should be encouraged as a further means of adding information to the flow matrix.

### **3.5 Modeling Annual Tons, Annual Shipment Value and Annual Ton-Miles**

Besides the 4 principal dimensions of the FAF O-D-C-M flow matrix, and the two added dimensions described above (i.e. the alternative data model and shipment distance interval dimensions), the operational flow estimation model also contains a seventh dimension. This dimension covers and relates the three different measures of annual freight activity we are simulating: annual tons shipped, annual dollar value of shipments, and annual ton-miles of freight transportation activity. While each of these measures can be treated within the current modeling framework as independent estimates of activity, and an O-D-C-D-M matrix generated for each, the FAF2 methodology also allows them to also be related through the log-linear model and IPF stages, as needed. The value of this sort of connection remains to be explored.

## **4. Example Application to the FAF Four Dimensional O-D-C-M Matrix**

In this section the *preliminary* results of an initial and key step in the above flow matrix building process are presented: generating the effects matrix from CFS Table 17. Doing so also highlights the data gap-filling challenge faced. The alternative data model and distance interval dimensions are not used in this example, for brevity. Table 17 is the most disaggregate of the published CFS tables, providing estimates of the annual tons, ton-miles and dollar value of shipments in the 43 SCTG 2-digit commodity classes shown in Table 1 (excluding Other/Unknown), also broken down by the 7 the modes of transport shown in Table 1. Note that Origins and Destinations in this table are by State, not FAF region. Working with these 50 State plus the District of Columbia regions for the present, the following describes selected effects (i.e. model parameter values) produced by a fully saturated 4-dimensional log-linear model of the form:

---

<sup>8</sup> Only four distance intervals are used to generate the initial FAF matrix, due to a temporary limitation on computer memory: 0-50, 50-100, 100-250, and > 250 miles.

$$F_{ijm} = \tau_0 \times \tau_i^O \times \tau_j^D \times \tau_m^M \times \tau_m^C \times \tau_{im}^{OM} \times \tau_{jm}^{DM} \times \tau_{cm}^{CM} \times \tau_{ij}^{OD} \times \tau_{ic}^{OC} \times \tau_{jc}^{DC} \times \tau_{ijm}^{ODM} \times \tau_{ijc}^{ODC} \times \tau_{icm}^{OCM} \times \tau_{jcm}^{DCM} \times \tau_{ijcm}^{ODCM} \quad (6)$$

### Single Dimension Effects

Table 2 shows, respectively, the individual origin state, destination state, commodity class and mode effects. The grand mean,  $\tau_0$ , in this run of the model = 4.67.

Table 2. One Dimensional Origin, Destination, Commodity and Mode Effects

#### Origins = States Plus DC, in alphabetic order

1.007	3.317	0.97	1.037	2.582	1.236	0.795	0.201	0.425	0.761
1.366	1.229	0.993	0.852	1.371	1.083	1.648	0.471	1.194	0.958
0.895	0.851	1.161	1.298	1.102	1.194	0.944	0.74	0.851	0.675
0.856	1.126	1.273	0.771	0.824	2.351	1.048	2.165	1.058	0.269
1.063	1.771	2.213	1.395	0.633	0.535	1.425	1.408	0.282	0.959
1.256									

#### Destinations = States Plus DC, in alphabetic order

1.272	1.756	0.792	1.174	2.894	0.959	0.824	0.453	0.276	1.966
1.395	2.988	0.886	2.562	1.316	0.809	1.006	0.913	0.846	0.454
1.007	1.194	1.881	1.03	0.641	1.139	0.917	0.78	0.919	0.679
1.743	0.575	1.455	1.61	0.704	1.719	0.754	0.725	0.977	0.709
0.841	0.477	1.237	2.22	0.697	0.479	1.061	0.686	1.426	1.094
0.709									

#### Commodities 1 through 43\*

0.231	1.968	2.344	2.779	0.952	1.34	2.402	3.653	1.496	0.229
0.693	2.04	0.535	3.659	7.85	0	7.593	5.746	2.868	1.14
1.336	2.071	0.615	0.397	0.148	0.844	0.731	0.68	0.233	0.186
0.693	0.782	0.27	0.469	0.753	1.017	0.369	0.426	0.364	0.334
1.095	2.551								

#### Other/

Truck	Rail	Water	Air	Truck-Rail	Other IM	Unknown
3.222	2.456	1.568	0.096	1.239	2.255	0.299

\* Commodity 42 = SCCTG 43 (Mixed Freight)

The following interpretations are in order. The larger- the-effect value, the greater its influence on the value of a cell's shipment volume. Hence large States such and California and Texas with comparatively large economies have  $\tau(i)$ s and  $\tau(j)$ s much greater than 1, and much greater than the average for all States. Smaller States such as Rhode Island and Delaware have, in direct contrast,  $\tau(i)$ s and  $\tau(j)$ s below 1.000.

The relationship between the  $\tau(i)$ s and  $\tau(j)$ s for each State is also of note, with a rough symmetry for most States but a clear asymmetry for a State such as Wyoming which exports a much greater tonnage, principally of Powder River Basin coal, relative to the annual tonnage of all freight coming into the State.

**Table 3. Mode-Commodity Effects  $\tau(m,c)$**   
**Coefficients: Value=0.9718 Tons=1.029**

<b>Modes (Rows):</b>	<b>Commodities (1 through 42*):</b>				
<b>Truck</b>	2.522	0.465	0.683	0.518	1.528
	1.02	1.305	0.947	0.178	0.824
	1.222	1.455	0.498	0.059	0.317
	0	1.834	1.121	0.887	0.696
	0.425	0.409	1.492	3.531	0.777
	1.647	1.08	1.272	2.218	2.297
	3.726	1.83	3.686	2.097	0.985
	1.156	0.421	0.741	1.238	2.614
	0.435	1.879			
<b>Rail</b>	0	0.759	2.505	0.589	0
	0.53	0.716	0.837	0	0
	1.788	0.587	2.043	1.517	1.322
	0	3.224	0.161	1.093	1.029
	0	1.86	0.499	2.016	0
	0.78	1.286	0.707	0	0
	1.859	1.322	4.896	0.508	0.087
	1.572	1.761	0	0	0
	0.781	0.737			
<b>Water</b>	0	4.866	4	5	
	0	2.548	0	0	0
	0.394	0.489	0	0	0.266
	0	1.271	2.154	8.456	7.523
	0	0	0	0.326	0
	0	0	0	0.181	0.568
	0.087	7.499	0.137	0	0
	0.142	0	0	0	0.169
	0	0			
<b>Air</b>	0.794	0	0.107	0	1.496
	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	2.195	0	0.668	0.539	0
	0	0	0	0.577	1.987
	0	0.444	0.412	0.994	2.152
	0.775	2.78	3.599	0	2.871
	0	0			
<b>Truck-Rail</b>	0	0	0.402	0	0
	1.766	0.588	0.272	0	0
	0.241	0	0	0	0
	0	0	0	0	0.64
	0	0	7.788	1.131	0
	0.614	0.342	2.209	2.219	0.639
	1.213	0.295	1.115	0.541	4.694
	4.154	0	0	0	0
	2.109	1.113			

**Table 3 (Continued)**

<b>Other Multiple Modes</b>	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	0	0.633	1.579	0	0
	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0
<b>Other &amp; Unknown</b>	0	0.546	0.934	1.902	0.433
	0.527	0.631	2.338	3.25	0
	4.264	2.246	0.525	5.977	8.414
	0	0.132	1.021	0.114	0.256
	1.509	0	0.316	1.521	0.744
	0.638	1.06	0.253	4.541	1.405
	1.209	0.903	1.895	2.135	1.421
	2.589	0.631	0.527	0.467	1.947
	0	0.326			

\* Commodity 42 = SCCTG 43 (Mixed Freight)

Similarly, high volume (tonnage) commodities such as grains also have high  $\lambda(c)$ s: while truck as a mode, which accounts for the preponderance of all tons moved annually also has a high  $\tau(m)$  relative to the other modes.

### *Example Two-Dimensional Effect*

Table 3 shows the mode-commodity cross effects. The effects in this table are shown as combined tonnage and value effects. The value and tons coefficients reported at the top of the table (i.e. 0.9718 for value and 1.0290 for tons), when multiplied through by the individual effects values in this table yield value only (in millions of dollars) and tons only effects (in thousands) respectively.

## **5. Imports and Exports**

Recalling the discussion in Section 2 above, the 2002 FAF flow matrix includes annual freight movements between the United States and seven foreign regions: our immediate neighbors Canada and Mexico, as well as Latin and South America, Asia, Europe, the Middle East, and Rest-of-the-World. Besides the CFS, the principal U.S. datasets dealing with imported and exported commodity shipments are USACE Foreign Waterborne Trade Data,<sup>9</sup> BTS Transborder Freight dataset,<sup>10</sup> and T-100 International Air Freight data.<sup>11</sup> This section provides a brief description of how these data were used in construction of the complete FAF flow matrix.

<sup>9</sup> See <http://www.iwr.usace.army.mil/ndc/usforeign/index.htm>. The import and export data are found at: <http://www.iwr.usace.army.mil/ndc/db/foreign/data/>

<sup>10</sup> <http://www.bts.gov/transborder/>

<sup>11</sup> <http://www.transtats.bts.gov/>

In addition to the freight entering and exiting via the nation's largest seaports and airports, a good deal of this trade now enters and leaves through the 14 FAF gateway regions. This includes the seven FAF Gateways located along the U.S.-Canadian or U.S.-Mexican borders (cf. Figure 2 and Table 1). The data fusion task facing FAF<sup>2</sup> requires that these annually imported and exported flows reflect not only the origination and destination, but also, where relevant, the FAF Gateway region through which they pass. It was decided therefore to develop an origin region - border region - destination region triple for each international movement.

This presents a significant data fusion challenge, beginning with conversion to FAF<sup>2</sup>'s 43 SCTG commodity classes from the Harmonized Schedule (HS) of commodity codes used in the STB dataset, and from the PMS codes used to record foreign waterborne commerce by the USACE. Both the waterborne commerce and multimodal surface transborder datasets each contain sufficient geographic detail to allow flows to and from each of the seven FAF foreign regions to be linked to the domestic FAF region of initial entry/exit. What they do not provide directly is the geographic detail necessary to associate these flows with originating or terminating domestic FAF regions.

***Waterborne Exports:*** While the CFS does provide estimates of the interior origination points and modes used for U.S. exports, the level of detail provided by these CFS export tables is much less than required, and generally less well developed than the tables of domestic movements [4]. The procedure used works as follows. Annual waterborne exports for each U.S. seaport are first cumulated into their respective FAF region. These flows are then traced back from each of these FAF seaport regions to their originating (i.e. U.S. internal) FAF regions using data from the CFS export tables to identify the most likely origination point for these shipments. Export flow totals in this case are based on the more reliable USACE waterborne commerce data set.<sup>12</sup>

***Waterborne Imports:*** Currently missing entirely from our national freight account is waterborne import data that ties the FAF internal region of destination to the appropriate foreign origination region. No readily available public domain database exists tracing either the true interior destination of most U.S. imports or the mode(s) by which this freight gets there [4].<sup>13</sup> For the purposes of generating the initial FAF flows matrix imported waterborne flows are treated as either terminating in the FAF region containing the U.S. seaport of arrival, or moving inland by one or more modes of transportation. The split between these local (assumed truck) movements and inter-regional moves is based currently on the split observed in the overall mode and destination shares reported for these same commodities within the entire FAF matrix. Further modeling of this data gap is clearly warranted and additional information needs to be identified for this purpose, perhaps on a commodity- or region-specific basis.

---

<sup>12</sup> Some 30% of all CFS export records in 2002 had their U.S. seaport of debarkation estimated.

<sup>13</sup> The (Port Import Export Reporting Service (PIERS) data set ( <http://www.piers.com/default2.asp> ), a product based on U.S Customs data offers the best available data on this activity. However, the data is proprietary, and is known to contain problems caused by reporting of inland freight destinations as the location of the importing company, rather than the true destination for these goods.

**Landed Imports and Exports:** The Transborder Freight datasets provide O-D data for shipments between U.S. States, Mexican States and Canadian Provinces. This same data source also provides the annual volume of freight moved by truck, rail, pipeline, moved as mail, and moved in other modes through each U.S. Customs-operated border crossing. By fusing these international O-D and border-crossing specific datasets, a set of flows that captures the FAF foreign region—FAF border entry/exit region – FAF internal U.S. region triples is produced. This is done by:

- assigning reported O-D volumes (after handling a number of idiosyncrasies in the dataset) to their most likely border crossings within each peripheral FAF region, with assignments weighted by the volume of traffic through each of these crossing points, and for each of the modes involved, then
- carrying out a proportional allocation of these U.S. State-based import and export volumes to those FAF Metropolitan and “Rest-of-State” regions within each State, based on the overall level of domestic trading reported by the CFS for each commodity within each FAF region.

**Air Imports and Exports.** The international air-freight data provided by BTS is in the form of total annual revenue tons transported between U.S. originating and terminating airports. No commodity detail is made available. For FAF<sup>2</sup> purposes these data are currently combined with the T-100 domestic air movements between FAF regions. That is, only flights between U.S. airport pairs are currently being modeled within the FAF. These data are being handled within the “alternative data model” dimension of the full FAF matrix as described in Section 3.3 above. Future work should tie these data directly to each of the seven foreign FAF regions.

## **6. Model Validation**

A method for testing the accuracy of the above approach is under development and will be used to ensure that estimated values of the missing cells in the FAF flow matrix are both reasonable and supported by the available statistical evidence. Ideally an independent estimate of the flows that this approach is trying to reproduce would be available. However, with the exception of annual state-to-state coal shipments (tonnages), as compiled and reported by the Energy Information Administration (based on its survey of coal shippers), there is little other commodity or industrial sector specific O-D-C-M data to draw on. This is, of course, the cause of our problem in the first place. The other source of data for validation purposes are the nation’s three major carrier-based and mode specific surveys: the STB railcar waybills, the Corps of Engineers waterborne commerce, and the Office of Airline Information’s air freight data. If any or all of these databases are used to help fill in the FAF matrix, as described in Section 3 above, then their use in validating the results of the data modeling is obviously tainted. There is also an additional problem with doing this. In all cases these mode-specific datasets do not produce true O-to-D matrices: rather, they report station-to-station flows, ignoring the presence of often quite lengthy, typically truck drays on one or both ends of a shipment. Proper

comparisons between these flows and FAF flows will therefore require these drays to be accounted in some manner.

In light of this situation a method that successfully reproduces known (CFS reported) cell values suggests itself. This is a common approach in statistical testing, in which a sub-set of the observed data is used to test the modeling of rest of the data set. Using this approach, known flow values are randomly and/or selectively removed from the observed FAF matrix. The log-linear model is then used to generate the necessary multi-dimensional effects and the IPF routine is used to re-produce the necessary cell values. These individual cell values can then be compared to their observed values, using comparative statistics such as absolute or percentage differences. Broader tests on the overall similarity of observed versus model-altered matrices can also be carried out, using root mean square error or log-likelihood based pseudo r-squares. The assumption being made here is that if the approach can successfully reproduce these known flows then it will also produce reasonable estimates of flows in the case of missing cells.

Given that the majority of cells that are missing data are likely to represent some of the smaller inter-regional/modal flows (for sampling reasons), this size bias may need to be accounted for in selecting known cells to reproduce. However, it should be noted that there are many examples of known high volume flows that are also suppressed within the 2002 CFS (and often similarly within the 1997 and 1993 surveys): for example where a few shippers of very large annual volumes exist within cells that fail either because they produce a coefficient of variation greater than 50% (the cut-off for CFS reporting) or for shipper confidentiality reasons.

## **7. Model Extensions and Work in Progress**

Adding a temporal dimension offers a further extension to the above-described approach, allowing data from the 1997 and also the 1993 Commodity Flow Surveys to influence the result. Doing so in a statistically defensible manner has its challenges, however, especially so given the quite significant changes in such statistics as truck shipment distances over the past decade. Some definitional differences between the three surveys will also need to be accounted for [4].

Still to be integrated into the matrix generation process at the time of writing are the domestic waterborne commerce and domestic and international air freight datasets discussed. Both provide additional quality controls in the resulting FAF flow matrix through inclusion within the “alternative data model” dimension. They may also prove valuable as a means of adjusting aggregate marginal totals where the CFS appears to have under- or over- counted the level of activity. Further work on integrating the CFS exports data into the process is also warranted, as is the development of a more sophisticated U.S. imports model. All of these improvements, plus the results of other commodity specific flow models can now be readily introduced into the multi-dimensional matrix generation process described in Section 3.

## References

1. Mani, A. and Prozzi, J. (2004) *State-of-Practice in Freight Data: A Review of Available Freight Data in the U.S.* Project report 0-4713-P2. Center for Transportation Research. University of Texas at Austin, TX 78705.
2. Meyburg, A. H. and Mbwana, J.R. (eds) (2002) *Conference Synthesis: Data Needs in the Changing World of Logistics and Freight Transportation.* New York State Department of Transportation, Albany, NY.
3. Southworth, F. (2003) *A Preliminary Roadmap for the American Freight Data Program.* Report prepared for the Bureau of Transportation Statistics, U.S. Department of Transportation, Washington D.C. 20590.
4. Southworth, F. (2005) *Filling Gaps in the US Commodity Flow Picture: Using the CFS with Other Data Sources.* Resource Paper, U.S. Commodity Flow Survey Conference, Boston, MA. July 8-9, 2005. Transportation Research Board, Washington D.C.
5. Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics, 11*: 427-444.
6. Goodman, L. A. (1971). The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics, 13*: 33-61.
7. Wrigley, N. (1985) *Categorical Data Analysis for Geographers and Environmental Scientists.* Blackburn.
8. Agresti, A. (1990) *Categorical Data Analysis.* Wiley, New York
9. Willekens, F. J. (1983) Log-linear modeling of spatial interaction. *Papers of the Regional Science Association, 52*: 87-205.
10. Southworth, F. and Peterson, B.E. (1990) Disaggregation within national vehicle miles of travel and fuel use forecasts in the United States. Chapter 7, *Spatial Energy Analysis*, L. Lundqvist, L-G Mattsson and E.A. Eriksson (Eds).

## Technical Appendix: A Complete Numerical Example

### A.1 Problem Definition

This section of the paper provides a complete numerical example of the hybrid log-linear modeling procedure described above, and its use within an iterative proportional fitting (IPF) scheme, showing how different sources of data can be combined to help fill in a commodity flows matrix. It also demonstrates the flexibility of the approach by showing how three different estimates of missing cell (missing commodity flow) values can be generated from the same basic approach but under different assumptions about the value of using specific data sources.

The 4x4 incomplete origin-to-destination matrix of annual commodity flows follows:

Origins	Destinations			
	1	2	3	4
1	300		60	90
2	200	500	30	60
3			300	80
4	40	80	150	200

The three grey shaded cells are the missing values that will be estimated.

For discussion purposes we will assume here that these values represent some multiple of annual tons shipped. In addition to this information we also have reported estimates of the total flow of the commodity coming into and going out of each of the four regions. This is the sort of information we can obtain from the U.S. Commodity Flow Surveys, i.e. reasonably comprehensive data on region-specific commodity productions and attractions, but incomplete data on the flows between regions (or their breakdowns within modes or commodity classes). The following table shows these origin and destination flow totals, which in total sum to 2,500 tons.

Origins	Destinations				
	1	2	3	4	
1	300		60	90	600
2	200	500	30	60	790
3			300	80	640
4	40	80	150	200	470
	639	888	540	430	2500

In what follows, an alternative solution for this flow matrix will be developed by using a spatial interaction model that makes use of the reported marginal totals given above and additional data on trip distances or costs to generate a fully filled in flows matrix. Note that an alternative source of such a matrix might be a second data source, such as a matrix based on the railcar waybill sample, or a matrix based on Corps of Engineers waterborne commerce data. *The idea here is to generate a matrix that provides initial estimates of the missing flows using some form of prior intelligence or external data to complete the original CFS-based flows matrix.*

## A.2 A Spatial Interaction Model

The first step is to fit a spatial interaction model to this trip production and consumption (“trip end”) data. For simplicity a SIA model of the form is selected:

$$T(i,j) = O(i)*D(j)*F[c(i,j)]*A(i)*B(j) \quad (6)$$

where  $T(i,j)$  = the annual tons shipped from region  $i$  to region  $j$

$O(i)$  = the annual tonnage shipped out of region  $i$

$D(j)$  = the annual tonnage shipped into region  $j$

$c(i,j)$  = a measure of travel impedance (cost, distance); and

$F[c(i,j)] = c(i,j)^{-1.5}$

The  $A(i)$  and  $B(j)$  are “trip end” balancing factors, i.e.,

$$A(i) = 1/\sum_j \{B(j)*D(j)*F[c(i,j)]\} \quad \text{for all } i \quad (A1)$$

$$B(j) = 1/\sum_i \{A(i)*O(i)*F[c(i,j)]\} \quad \text{for all } j \quad (A2)$$

Iterating between these  $A(i)$ s and  $B(j)$ s ensures, respectively, that

$$\sum_j T(i,j) = O(i) \text{ for all } i \quad \text{and} \quad \sum_j T(i,j) = D(j) \text{ for all } j \quad (A3)$$

That is, we get back our reported trip production and attraction (i.e. origin and destination) flow totals.

Solving this model yields the following matrix of estimated flows:

<b>331</b>	<b>136</b>	<b>46</b>	<b>87</b>	<b>600</b>
<b>178</b>	<b>548</b>	<b>17</b>	<b>47</b>	<b>790</b>
<b>82</b>	<b>145</b>	<b>340</b>	<b>73</b>	<b>640</b>
<b>48</b>	<b>59</b>	<b>139</b>	<b>224</b>	<b>470</b>
<b>639</b>	<b>888</b>	<b>542</b>	<b>431</b>	<b>2500</b>

In practice, such an SIA model has to be calibrated to determine the travel impedance parameters (give here simply as -1.5) and perhaps additional factors of importance to a specific commodity’s flow pattern.

One way to use this interaction modeling result is to introduce its values for the three missing cells into the original matrix. This produces the following matrix:

Origins	Destinations				
	1	2	3	4	
1	300	136	60	90	586
2	200	500	30	60	790
3	82	145	300	80	607
4	40	80	150	200	470
	622	861	540	430	2453

A new row and column margins emerge. To recover the original margins, while still retaining the original values of all reported cells, an iterative proportion fitting can be used to adjust our three missing value estimates until the reported marginal totals are recovered. This can be done by first extracting all of the observed cell values from the reported row and column totals. This leaves the following marginal residuals:

					150
					0
					260
					0
	99	308	0	0	410

IPF is used to force the three missing valued cell estimates to conform to these residual row and column totals, giving the following result (rounded to nearest integers):

		149			150
					0
	99	159			260
					0
	99	308			410

Putting these numbers back into the original flows matrix then yields the final filled in matrix of flows, matched to the original row and column margins (after rounding).

Origins	Destinations				
	1	2	3	4	
1	300	149	60	90	600
2	200	500	30	60	790
3	99	159	300	80	640
4	40	80	150	200	470
	639	888	540	430	2500

This is one way to estimate the missing flows. A second and more demanding approach is presented below.

### A.3 Log-Linear Modeling of Missing Cell Values

Next a log-linear model that incorporates data from both the original flows matrix and this SIA model estimated matrix is created. This model has the form:

$$T(m,i,j) = \alpha * \tau(i) * \tau(j) * \tau(m) * \tau(i,m) * \tau(j,m) * \tau(i,j) * \tau(i,j,m) \quad (A4)$$

where  $T(i,j,m)$  = the annual tons of a commodity moved between origin location  $i$  and destination location  $j$ , according to data model “ $m$ ”. Here data model  $m$  = Model A or Model B, where Model A is the initial “observed” FAF matrix and Model B refers to the flows estimated by the spatial interaction model (or, if we wish, by an external data source). In practice this model in terms of its natural logs will be solved as:

$$\ln T_{ijm} = u + \lambda_i + \lambda_j + \lambda_m + \lambda_{im} + \lambda_{jm} + \lambda_{ij} + \lambda_{ijm} \quad (A5)$$

where,

$$u = 1/(I*J*M) * \sum_{i,j,m} \ln [T(i,j,m)] \quad (A6)$$

$$\lambda_i = 1/(J*M) * \sum_{j,m} \ln [T(i,j,m)] - u \quad \text{for all } i \quad (A7)$$

$$\lambda_j = 1/(I*M) * \sum_{i,m} \ln [T(i,j,m)] - u \quad \text{for all } j \quad (A8)$$

$$\lambda_m = 1/(I*J) * \sum_{i,j} \ln [T(i,j,m)] - u \quad \text{for all } m \quad (A9)$$

$$\lambda_{i,j} = 1/(M) * \sum_m \ln [T(i,j,m)] - (u + \lambda_i + \lambda_j) \quad \text{for all } (i,j) \text{ pairs} \quad (A10)$$

$$\lambda_{i,m} = 1/(J) * \sum_j \ln [T(i,j,m)] - (u + \lambda_i + \lambda_m) \quad \text{for all } (i,m) \text{ pairs} \quad (A11)$$

$$\lambda_{j,m} = 1/(I) * \sum_i \ln [T(i,j,m)] - (u + \lambda_j + \lambda_m) \quad \text{for all } (j,m) \text{ pairs} \quad (A12)$$

$$\lambda_{i,j,m} = \ln [T(i,j,m)] - (u + \lambda_i + \lambda_j + \lambda_m + \lambda_{i,j} + \lambda_{i,m} + \lambda_{j,m}) \quad \text{for } (i,j,m) \text{ triplets} \quad (A13)$$

for  $i=1\dots I$  ( $=4$ ) origin regions,  $J=1\dots J$  ( $=4$ ) destinations, and  $m=1\dots M$  ( $=2$ ) data models.

Equation (10) is derived by solving for  $\alpha$  and for each of the  $\tau$  terms. In log-linear parlance  $\alpha$  is referred to as the grand mean and each  $\tau$  as the “effect” on the resulting flow estimate from a specific dimension of the problem: analogous to the effects in an analysis of variance. That is, there are three one-way effects, (( $i$ ), ( $j$ ) and ( $m$ ), three two-way effects ( $i,j$ ), ( $i,m$ ) and ( $j,m$ ), and one three-way effect ( $i,j,m$ ). For example  $\tau(i,j)$  = the contribution to the value of  $T(i,j,m)$  interaction from the pattern of interactions between  $i$  and  $j$ , irrespective of the mode of transport used. Equation (5) is now applied to the following  $4(i) \times 4(j) \times 2(m)$  matrix of flows:

Model A (Reported) Flows:

	300		60	90
	200	500	30	60
			300	80
	40	80	150	200

Model B (SIA Model) Flows:

331	136	46	87
178	548	17	47
82	145	340	73
48	59	139	224

Solving equation (11) for this matrix produces the following result:

300	275	60	90	725
200	500	30	60	790
125	251	300	80	756
40	80	150	200	470
665	1106	540	430	2741
331	136	46	87	600
178	548	17	47	790
82	145	340	73	640
48	59	139	224	470
639	888	542	431	2500

The missing cells in the original Model A matrix now contain values. The Yellow (shaded) cells on the margins of this and the Model B matrix here are simply the row and column summations, i.e. the  $O(i)$ s and  $D(j)$ s. For example, the number 725 in the top right-hand corner of this matrix represents the sum of all flow out of origin (row)  $i=1$  according to Model A, after initial gap filling (i.e.  $300+275+60+60=725$ ). Assuming that the SIA model has been generated using the observed set of  $O(i)$ s and  $D(j)$ s for this problem, then the cells in the Model A matrix need to be readjusted to match these “observed” totals. (Note that rounding error is the only reasons columns 3 and 4 differ between the two models in this example). This is accomplished by using IPF.

#### A.4 Results of Iterative Proportional Fitting on the Gap Filled Flow Matrix

The table below shows the results from the first six of these iterations using IPF to reconcile the flows in the matrix, first to the row and then to the column totals. After six iterations, the result comes very close to the observed set of  $O(i)$ s and  $D(j)$ s (and could be forced to fit exactly). In the process the values in the three missing cells have been reduced so that they come closer to SIA model result, while retaining the higher values implied for these by the reported data (Model A). In doing this, the value for a number of Model A reported cells have been altered, causing some of them to better approximate the SIA model result also. The final result then is a compromise between reported and modeled flow data that also matches all reported marginal totals, estimates missing cell values, and does so while disturbing the structure of the rest of the flows matrix as little as possible.

248	227	50	75	600
200	500	30	60	790
105	213	254	68	640
40	80	150	200	470
594	1020	484	402	2500

267	198	56	80	601
215	435	34	64	748
113	185	285	73	656
43	70	168	214	495
639	888	542	431	2500

267	198	56	80	600
227	459	35	68	790
111	181	278	71	640
41	66	160	203	470
646	904	528	422	2500

264	194	57	81	597
225	451	36	69	782
110	178	285	72	644
40	65	164	208	477
639	888	542	431	2500

266	195	57	82	600
227	456	37	70	790
109	176	283	72	640
40	64	161	205	470
641	892	538	429	2500

265	194	58	82	599
226	454	37	70	788
108	176	285	72	641
40	64	162	206	472
639	888	542	431	2500

Of note in this example are the rather large values estimated for the three missing cells. This results from their proximity to large diagonal flows in the reported, “Model A” flows matrix. These adjacent intra-zonal flows appear to have too much of an effect on the missing cell estimates (IF we believe the SIA Model B, results, that is). Such effects are common to spatial interaction matrices. One way to handle this is to solve the log-linear model and IPF on inter-zonal flows only. Doing so in this case produces the following initial estimates for the three missing cells:

	86	60	90	236
200		30	60	290
80	89		80	249
40	80	150		270
320	256	240	230	1045
	136	46	87	269
178		17	47	242
82	145		73	300
48	59	139		246
308	340	202	207	1057

that are now much lower than the Model B values in each case. Performing IPF on this modified Model A matrix yields the following result (after six iterations):

0	122	60	87	269
176	0	23	43	242
95	128	0	77	300
37	90	119	0	246
308	340	202	207	1057

The missing cell values are now much closer to their SIA model estimated values, i.e. 122 vs. 136; 95 vs. 82; and 128 vs. 145.

### A.5 Discussion of the Results

The above analysis has provided us with three different estimates of the missing value cells, as follows:

	SIA model w/IPF	Log-Linear+SIA model w/IPF	Log-Linear+SIA model w/IPF, w/o Intra-zonals
cell 1,2	149	194	122
cell 3,1	99	108	95
cell 3,2	159	176	128

The combined Log-Linear with SIA Model approach offers one method for trying to make the most use of the existing data on commodity flows, while recognizing that some form of intelligent modeling is needed to fill in the values for missing cells. In this case this intelligence is introduced largely through a spatial interaction model that recognizes a common pattern in commodity movements: that the volume of such flows tends to attenuate with extra cost or distance moved. The SIA model might also be used more directly to estimate missing cells, using IPF to adjust these cell estimates to fit reported row and column totals. However, experience with fitting such interaction models indicates that there are often other factors at work that often cause such models, unless developed in rather elaborate frameworks using additional variables, to miss some of the real-world unevenness in the data.

If a flow matrix has most of its cells reported, of course, then the need for elaborate modeling is limited, at least for base year estimation. A simple interaction model or even a simple IPF routine might be used to fill in missing cells and match the result to observed, marginal totals. At the other extreme, where very few if any cells of the flow matrix are filled in, heavy reliance on some form of the more elaborate interaction model for the flow estimates is needed. Between these two extremes, sufficient reported data on some flows may be available to warrant their inclusion, and possibly retention of their reported value, in the matrix filling process. This is a gray area, however. If the above-described log-linear modeling technique is applied, then these cell values will change to some degree to accommodate missing value insertions, subject to marginal totals. If confidence intervals on cell values are available, then a comparison might be made of the amount that these values change from their original reported values and acceptance given where the value remains within the selected confidence interval. Marginal totals can also be allowed to vary and be tested in a similar fashion. In the above example it is assumed that these  $O(i)$ s and  $D(j)$ s are more robust estimates than individual cell values, and hence they are worth retaining. This isn't always true. However, the cost of allowing some latitude in the value of marginal totals further complicates the problem and makes comparison to "official" totals something of a headache. In the final analysis the choice of modeling approach is an empirical one and depends on how believable a spatial interaction or other form of model is reproducing the real world flow pattern.