1. Introduction

The purpose of this paper is to describe the task of constructing a set of commodity class and mode specific annual origin-to-destination flows for the entire United States, covering all domestic as well as all imported and exported goods. The project is funded by the Office of Operations, Freight Management and Operations, within the Federal Highway Administration, as part of the U.S. Department of Transportation’s Freight Analysis Framework (FAF3) program. The task takes as its starting point the 2007 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 and annual dollar values for the commodities moved, with the principal dimensions being:

- 123 domestic plus 8 foreign shipment origination regions (O)
- 123 domestic plus 8 foreign shipment destination regions (D)
- 43 classes of commodities being transported (C), and
- 7 modes of transportation used (M), including intermodal combinations

Having identified the problem to be solved, the paper describes how log-linear modeling can be used to estimate the values for many of the zero valued cells found in large, sparse matrices constructed from sample surveys, 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 a variety of reported marginal flow totals. Results are presented for portions of the full flow matrix. The paper also describes how the approach can be extended to a) incorporate a temporal dimension covering past CFS databases, and b) how it is being extended to draw in data from other mode and commodity specific data sources where deemed useful to the estimation/data gap filling process. A third extension, using shipment distance interval data as one more dimension to iterate on, is also discussed.
2. Nature of the Missing Data CFS 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 question we must answer is not only what size each of these flows should be, but also, which cells ought to contain a positive flow at all. It should also be noted that even when we sum the data across a particular row or column in the CFS O-D-C-M matrix we sometimes find that we are missing data for these two as well as three dimensional margins: and not just data on the four-dimensional flows we are seeking. Study of the complete set of 2002 CFS data products indicates that we have a good many data matrices to work with. This includes the most detailed of the published matrices, which reports annual tons, dollar value and ton-miles shipped by state of origin, state of destination, mode and 2-digit (43) commodity classes.\(^1\) Other tables provide 1, 2 and 3 dimensional looks at this same data, including flows broken down to the 123 CFS/FAF intra-US geographic regions of interest. Without going through the contents of each data table in turn, 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.

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

3. 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

\(^1\) [http://www.census.gov/svsd/www/02CFSdata.html](http://www.census.gov/svsd/www/02CFSdata.html)
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

While 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 paper also describes a method for expanding the FAF3 flow matrix creation problem by three additional dimensions:

4. The Matrix Generation Methodology

4.1 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 [1,2]. 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:

\[
T(i,j, r+1) = \frac{T(i,j, r)}{\sum_j T(i,j, r)} \times O(i) \tag{1}
\]

\[
T(i,j, r+2) = \frac{T(i,j, r+1)}{\sum_i T(i,j, r+1)} \times D(j) \tag{2}
\]

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 we get:

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

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.

4.2 Log-Linear Modeling of Missing Flows

Numerous practical examples of applying IPF-based methods exist; including some
directly relevant applications to multi-dimensional movement tables [3,4,5,6]. The FAF3
log-linear freight flows model has the following “fully saturated” form:

\[
F_{ijcm} = \lambda_0 \times \lambda_i^O \times \lambda_j^D \times \lambda_m^M \times \lambda_{im}^C \times \lambda_{jm}^{OM} \times \lambda_{jm}^{DM} \times \lambda_{ij}^{CM} \times \lambda_{ij}^{OD} \times \lambda_{ic}^{OC} \times \lambda_{jc}^{DC} \\
\times \lambda_{ijm}^{ODM} \times \lambda_{ijc}^{ODC} \times \lambda_{ijcm}^{OCM} \times \lambda_{ijcm}^{DCM} \times \lambda_{ijcm}^{ODCM}
\]

which is solved computationally using natural logs, i.e.

\[
\ln(F_{ijcm}) = \lambda_0 + \lambda_i^O + \lambda_j^D + \lambda_m^M + \lambda_{im}^C + \lambda_{jm}^{OM} + \lambda_{jm}^{DM} + \lambda_{ij}^{CM} + \lambda_{ij}^{OD} + \lambda_{ic}^{OC} + \\
\lambda_{ijc}^{DC} + \lambda_{ijm}^{ODM} + \lambda_{ijc}^{ODC} + \lambda_{ijcm}^{OCM} + \lambda_{ijcm}^{DCM} + \lambda_{ijcm}^{ODCM}
\]

where \(F_{ijcm}\) = the annual volume of commodity ‘c’ moved, by mode ‘m’ (in tons, or
dollars) between FAF3 origin zone ‘i’ and FAF3 destination zone ‘j’ in 2007, and where
the \(\lambda\)’s represent the log-linear model parameters to be estimated, and often termed
the “effects” of the different dimensions, or combinations of dimensions, on this result. For
example, \(\lambda_{im}^{OM}\) represents the effect of a specific shipment origin i and mode
combination m on the resulting freight flow estimate. Given a completely filled in flows
matrix equation (5) will reproduce the cell estimates exactly. We are interested for FAF
purposes in how such a model performs with missing data.

4.3 Handling Zero-Valued Cells

The FAF3 flow matrix is a very large and very sparse multi-dimensional table. That is, it
contains a large number of zero valued cells. We need to determine which of these cells
are true or structural zeros, and which are zero valued because of the limitations of CFS
sampling. We term these latter sampling zeros. In the CFS the published tables there are
also many cells containing the letter “S”: signifying a cell estimate too poor in its
statistical properties to be reported (including all cells with a coefficient of variation of
greater than 50%), (use of the letter “D” signifying a need to suppress the cell value to
avoid disclosing data about individual company activity is also possible). While such
suppression causes us to lose data it also provides information that we can use in
generating the cell-specific effects we are seeking. That is, where an “S” (or a “D”)
occurs we know that there exists a positive flow value for that cell. The log-linear model
is used to generate a suitable effect, and subsequently a positive flow estimate, for such a
cell.
But how good is our 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 we have only small O-D flows to deal 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 our ability to both identify and estimate these missing cell values additional data modeling steps have been taken, and these will be described.

5. Expanding the Methodology to Capture Additional Dimensions of Freight Movement

5.1 Adding an “Alternative Data Model” Dimension

The limited sample sizes of past U.S. commodity Flow surveys mean that other freight data sources can often improve the estimation process if a method can be found to merge them in a statistically acceptable way. This applies in particular to 100% carrier based surveys of U.S. waterborne, rail, and airborne freight movements collected by the US Army Corps of Engineers, Surface Transportation Board, and Bureau of Transportation statistics respectively. Each of these datasets can provide additional quality controls on the resulting FAF multimodal freight flow matrix, through inclusion in the LLM//IFP process via an “alternative data model” dimension. They can also prove valuable as a means of adjusting aggregate marginal totals where the CFS appears to have under- or over- counted the level of freight activity. The method used when creating the 2007 FAF3 multi-dimensional freight flows matrix is described

5.2 Adding a Temporal Dimension

Adding a temporal dimension offers a further extension to the approach, allowing data from the 2002, 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 [7].

5.3 Adding a Shipment Length Dimension

The pros and cons of using the shipment distance data reported by the US Commodity Flow Survey as one more set of marginal control totals will also be discussed. This translates the resulting flows into a set of a linked set of spatial interaction models [5]. The pros and cons of doing so are discussed.
References


