Filling Gaps in the U.S. Commodity Flow Picture: Using the CFS with Other Data Sources

by
Frank Southworth
Oak Ridge National Laboratory
Oak Ridge, TN 38731
sou@ornl.gov

June 8, 2005

Prepared for
The 2005 Commodity Flow Survey
Users' Conference
Boston, MA.
July 8-9, 2005

Prepared by the
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831-6073
managed by
UT-BATTELLE, LLC
for the
U.S. DEPARTMENT OF ENERGY
under Contract No. DE-AC05-00OR22725

Abstract

The U.S. Commodity Flow Surveys (CFS) provide a wealth of data on the movement of freight within the country, by mode of transportation, dollar value, annual tonnage, door-to-door shipment distance and shipment size. Yet gaps in CFS coverage and a lack of spatial and commodity detail limit the value of this data for planning and policy supporting studies. This paper identifies these gaps and considers how supplementary data sources combined with statistical modeling techniques can be used to create a more complete picture of national and regional commodity flows. The first half of the paper describes current data gaps and available data resources. The second half of the paper discusses possible gap-filling solutions. Solutions fall into two categories: data synthesis using current data sources, and improved and expanded data collection methods.

1. Introduction

Following a sixteen year hiatus in federal data collection the 1993, 1997 and 2002 U.S. Commodity Flow Surveys (CFS) have filled a large gap in the U.S. freight data universe. As a result of these surveys we now have data on the annual volume of commodity movements taking place into, out of, and within each of our States, the District of Columbia, and our largest metropolitan areas, broken down by mode of transport. We also have data on the "door-to-door" travel distances and shipment sizes associated with these movements, as well as both their annual tonnages and dollar value. However, in trying to use these data in planning and policy supporting studies we run into two kinds of problems. First, the surveys do not cover all U.S. freight movements. Second, the surveys only support the representation of origin-to-destination (O-D) movements between quite large geographic regions, and even then the O-Ds they do produce are limited in the level of commodity detail the surveys can support (due to sample size), or are allowed by law to reveal.

Gaps in coverage, difficulties in determining where different data sets overlap, and the need to consult a number of different data dictionaries in order to develop an adequate representation of a region's freight movements can prove more than just frustrating. They also make it difficult to establish with confidence the accuracy and statistical robustness of the resulting estimates of these commodity movements. As recent reviews have pointed out, this is the data universe the freight analyst is faced with currently, whether dealing with nationwide, statewide, or metropolitan area-wide commodity flows [1-4].

The purpose of this paper is to describe succinctly these gaps in CFS coverage and to discuss ways that other data sources can be used to fill them. The principal beneficiaries of a successful gap filling activity are analysts and policy makers in federal, state and metropolitan agencies where the volumes of goods moved are key inputs to investment decisions affecting future transportation infrastructures and services. In particular, the paper is used to explore the following questions:

1. What other data sets are available for filling the current gaps in CFS coverage and detail? (Section 2)
2. What data modeling techniques exist for combining CFS data with these other data sources? (Sections 3)
3. What new and emerging ways of collecting freight movement data might we tap into? (Section 4)

Section 5 of the paper provides a summary of questions raised and directions for future research and development.

The data products we need. Before entering this discussion it's useful to define the nature of the data products we're looking for. For present purposes two data products are of most interest, whether developed on a fully national or individual statewide basis:

- a multi-dimensional freight flow matrix, the principal dimensions of which are freight traffic generating origins, destinations, modes and types of commodities moved; and based on this flows matrix:
- a series of traffic assignments showing how freight vehicles move over a region's roadways (highways, railways, waterways, pipelines) and through its various modal (e.g. truck village) and intermodal (seaport, airport, truck-water, truck-rail and truck-pipeline) transfer terminals.

With these two products we can address a wide range of issues and studies commonly faced by regional planners: from the creation of time-series statistics on total freight activity, to an analysis of competition and cooperation between modes, to an analysis of the economic, safety, and environmental impacts of site-specific transportation capacity expansion or service modification projects. Both products, of course, imply the availability of other supporting datasets, notably data to represent a transportation network, as well as coefficients that translate emissions, fuel, and travel times into suitable movement costs. Focusing on the uses to which the CFS can be put, it is with the quality of the supporting "freight movement data" that we are concerned here.

2. The Data Challenge: What the Data Do and Don't Tell Us

For the purpose of generating detailed freight movement matrices, the O-D data sets available to most users can be grouped into three classes:

1. A nationwide multimodal commodity shipment survey of "door-to-door" movements (i.e. the U.S. Commodity Flow Surveys)
2. Nationwide and mode specific freight carrier activity surveys reporting "station-to-station" freight movements, reporting tons moved by specific industries and types of transportation equipment
3. International trade and trans-border traffic flow surveys, reporting principally commodity specific, dollar valued trades.

Taken together these data represent a loosely connected patchwork quilt with a number of holes in its coverage. The reader is directed to references [1] through [5] for recent listings and discussions of currently available databases falling under each of these headings. The principal gaps in each type of data are summarized below.

2.1 CFS Strengths and Weaknesses

The CFS has a number of unique strengths. In particular:

- It is fully national in scope
- It covers all the major surface transportation modes (truck, rail, water, petroleum pipelines), as well as shipments of air freight;
- It identifies the true geographic origin and destination of each shipment (and therefore also provides estimates of "door-to-door" shipment distances);
- It collects data on both the weight and dollar value of all in-scope shipments; and
- It has a time series in the form of the 1993, 1997 and 2002 surveys
- It is done in conjunction with the Economic Census, providing concurrency with other datasets.

On the debit side, in particular:

- Not all commodities are covered by the CFS
- The survey does not, in theory, capture imports
- The spatial detail available to its mode specific O-D matrices is limited to a small number of rather large geographic regions
- The volume of "intermodal"[1] freight reported may be low, due at least in part to definitional issues
- The shipment length detail available from non-geographically disaggregated products is very limited in its supporting commodity-level detail
- The surveys have seen some content changes, and a 4 to 1 reduction in sample size between 1993 and 2002 that makes for some large coefficients of variation in reported estimates.
- There are discrepancies in the estimates generated by the CFS and the U.S. Army Corps of Engineers' waterborne commerce data, the latter based on industry-wide carrier reporting that produces larger ton and ton-mileage figures.

Coverage Issues: Commodities. All three (1993, 1997, 2002) CFS surveys sampled business establishments in mining, manufacturing, wholesale trade, and selected retail industries [6]. The surveys also cover selected auxiliary establishments, such as the warehouses of in-scope multi-unit and retail companies. The surveys do not cover establishments classified as farms, forestry, fisheries, construction, transportation (including household goods carriers), governments (including military and mail shipments), foreign establishments, services, and most establishments in retail. As a result, the CFS has been conservatively estimated to cover

[1] i.e. shipments requiring end-on transfers between two different modes in order to reach their destination

less than 75% of all the freight tons moved annually in the United States [7]. A common trait of many of the missing shipments is the dominance of highly localized, essentially truck-only movements. This applies to most retail, and also to a good deal of activity in the construction and personal delivery services industries, both significant sources of short range truck miles of travel according to the 1997 and 2002 U.S. Vehicle Inventory and Use Surveys [8].

Coverage Issues: Imports. Being a survey of U.S. shippers the CFS also does not capture imports. Here a difficulty in adding imported tonnages to the CFS arises because a shipment that may be classified as imported cargo in the international trading arena also finds itself being re-defined as an internal shipment within the CFS once a change in ownership of the goods has taken place. This may occur at a warehouse or other storage facility located close to the U.S. port of debarkation. How often this occurs is unclear. The CFS does include goods exported by U.S. shippers. However, the estimated volume and value of these shipments is often at odds with the reporting of such exports in the U.S Foreign Trade data.

O-D Detail. Perhaps the single most significant improvement to the CFS for planning purposes would be the addition of geographic detail. This is asking a good deal. Even for the current 114 CFS regions, once shipments are broken down by both origination and destination region the level of both commodity and modal detail that the Census Bureau can release becomes limited, due to lack of robustness in sample-based estimates, or to the need to protect the confidential nature of shipper activities. These constraints restrict the survey's O-D matrices to very general 1- or 2-digit commodity classes. At the 2-digit level this represents only 42 commodity groupings. And not even the 1993 survey, based on a sample of some 200,000 shipping establishments, was able to produce complete commodity and mode specific O-D matrices at the level of rather broad 2-digit commodity codes. With the 1997 CFS sample reduced to some 100,000 establishments, and with the 2002 CFS further reduced to a sample of 50,000 shippers, data on the origin-destination-commodity-mode (ODCM) combinations most useful to transportation planning agencies proves quite sparse.

Intermodal Shipments. The survey also faces a difficult task in representing the volume of intermodal freight movements, i.e. movements of freight from origin to destination using two or more end-on modes. One reason for this is probably definitional, and related to ownership of cargo. Another may be that some survey respondents didn't know how their product reached its final destination, only the mode it left their establishment in. This is, anyway, a generic problem that has to be faced when tracking any product through its freight movement supply chain. For example, a grain shipment will typically change hands when it reaches a grain elevator. The farmer sells to the elevator, who in turn sells to a customer "down river". In contrast, a coal shipment may belong to the mine owner all the way to the utility at which it is consumed, possibly via truck-rail, truck-water or rail-water transfer. Parcel shipments are also recognized as a separate category of freight in the CFS, and these too probably involve more intermodal activity than is captured.

Shipment Lengths. To date the CFS program has provided limited additional assistance to the spatial analyst. In particular, by limiting distance reporting to a small number of rather broad distance intervals, and for rather broad aggregations within commodity classes, a good deal of information that could be used to inform spatial interaction modeling is not available. Out-of-scope CFS truck and rail freight movements, notably of imported goods, pose a similar problem, with the Transborder Surface Freight Data and the Port Import and Export Reporting Service datasets each providing their respective US-Canadian, US –Mexican and US-transoceanic movements at only the State level (and often for a destination address associated with the business office of the receiver, rather than with the true destination of cargo delivery).

Time-Series Issues. As a time series of freight movement activity, the reduction in CFS sample size from roughly 200,000 establishments in 1993 and to 50,000 in 2002 is of great concern. As a result, some of the more detailed O-D matrix elements available in 1993 are no longer statistically robust enough to be reported in 2002. The CFS has also seen a number of changes in its design and content that can affect trend analyses. One of these changes was the move from the original system of 89 National Transportation Analysis Regions (NTARS), to a 1997 regionalization based around the nation's 56 most populous metropolitan areas, small states, and remainder-of-state regions. The 1993 CFS used STCC commodity codes, the 1997 and 2002 surveys moved to SCTG. Between 1997 and 2002 the boundaries of some of these metropolitan areas also changed. Of more concern was the impact of the NAICS re-classification of the underlying business establishments from which the survey was drawn. This resulted in a loss of data for both the lumber and printed matter industries, as covered in the 1993 and 1997 surveys (i.e. they fell out of scope). In 1997 the request for data on containerized shipments was dropped, while the method of asking for information on hazardous materials shipments was improved, ensuring among other things that petroleum shipments were captured in this category. The difficulty of creating petroleum pipeline O-Ds from the 1993 survey data led to this aspect of the survey being down-played in the 1997 and 2002 data creation efforts. Some of these changes may prove important when trying to construct temporal trends in ton, dollar, ton-mile or vehicle-mile statistics; or if trying to use the 1993 and 1997 datasets to fill gaps in the 2002 data.

If such trend information is important then we might also ask whether a different approach to the survey design is in order. In particular, would a continuously sampled CFS help? This last issue is discussed further in [4] where the potential strengths as well as the practical challenges of continuous sampling are outlined, with evidence for some success in past freight surveys.

2.2 The Mode Specific Carrier Surveys

In contrast to the CFS, the nation's carrier-based surveys are mode specific. Given the CFS problems of scope and detail discussed above, these surveys offer a natural option for enhancing the freight data picture.

True Versus Line-Haul Os and Ds. Via the Surface Transportation Board's Rail Carload Waybill sample[2] and data contained in the Army Corps of Engineers' Waterborne Commerce database[3] it is possible to estimate, respectively, the volume of freight moving over specific sections of a railroad's track and over specific reaches of the nation's navigable waterways, and with a little work to associate these shipments with specific station-to-station or dock-to-

[2] http://www.stb.dot.gov/stb/industry/econ_waybill.html
[3] http://www.iwr.usace.army.mil/ndc/

dock routings over the U.S. rail and water networks. Both of these annual data collection efforts cover all of the commodities moved by their respective modes as well as details of the types of railcars and barges used to move different types of freight. The waterborne commerce data is an approximately 100% sample of inland barge, intra-coastal and Great Lakes movements. The rail waybills are a much smaller sample, but emphasizing the larger and heavier unit train movements. (They do not capture export shipments carried on the Canadian railroads operating inside the U.S., however).

What is missing for the purpose of O-D estimation is the true origin and destination of these movements. This poses a problem, since a significant amount of the freight transported by rail and water involves truck draying to or from a rail or barge terminal, sometimes involving travel distances that place either or both the origin and destination of the freight outside the CFS region or State associated with the railcar or barge line-haul movement. This lack of true O and D information also applies to our current surveys of air freight carriers, again requiring that we somehow infer the true origin and destination of the truck drays that are involved in the vast majority of these low weight but high valued shipments. The most readily accessible forms of this air freight data, the Office of Airline Information's Air Freight Statistics, also provides only total tons of freight (and mail) transported, without commodity breakdowns.

More Data Needed on Trucks. The principal source of data on U.S. truck activity, the Vehicle Inventory and Use Surveys (VIUS)[4] contains no O-D data per se. This makes the CFS the only source of nationwide data on O-D truck movements, and the CFS captures no details on the type of truck used other than its for-hire versus private ownership status. What the VIUS offers is considerable detail, and a time series back to 1963, on the types of trucks used to haul freight of different types. In doing so it has a number of strengths and weaknesses of its own. On the credit side, it offers operator estimates of each vehicle's annual activity broken down by commodity carried and typical operating range (in distance intervals). The vehicle characteristics data is especially rich, including data on a vehicle's body type, length, axle configuration, empty and loaded operating weigh and mileage, ownership, fuel use, and hazardous cargo transport. On the debit side the commodity detail is quite limited: 33 classes prior to 2002, expanded in the 2002 VIUS to 51 classes based on the 2-digit SCTG codes used by the CFS. The 5 year interval between each VIUS also means that we often have to wait a few years to see the effects of any important changes in trucking practices.

Too Many Commodity Classification Schemes. Direct comparison to, or combination of, information from the above carrier surveys with CFS data, or with each other, is further complicated because each has its own commodity coding scheme. While in 1997 and 2002 the CFS commodities are classified according to the Standard Classification of Transported Goods (SCTG), the rail waybills use STCC (Standard Transportation commodity Codes), and the waterborne commerce data is based on yet another commodity classification scheme. U.S. Foreign Trade Statistics are based on yet a fourth scheme: the Harmonized Commodity and Coding System, and the United Nations uses yet another scheme (the Standard International Trade Classification, or SITC) for reporting international trades. The VIUS, which previously used SIC (Standard Industrial Classification) commodity/industry codes, was converted to

[4] http://www.census.gov/svsd/www/products.html

SCTG codes in the 2002 survey, making it compatible with the CFS. Finally, the coding scheme associated with the economic activity data sets commonly used to support both forecasting and geographic disaggregation of CFS-based commodity flows (see Section 3 below) is the Census Bureau's North American Industrial Classification System, or NAICS (which replaced the SIC codes used by the Economic Census prior to 2000).

To combine data from two or more of these data sources means using a suitable "cross-walk" between the different commodity/industrial sector coding schemes. While such cross-walks already exist, there is necessarily some degree of lost accuracy in the resulting merger, especially at the more aggregate levels of some of the more diversified commodity grouping.

2.3 U.S. Trade Data: Movements of Imported and Exported Cargo

A number of federal agencies are involved in the collection, processing and dissemination of international trade and transportation data. U.S. merchandise trade statistics are processed and released by the Foreign Trade Division of the U.S. Census Bureau. International merchandise trade data are captured from administrative documents required by the Departments of Commerce and Treasury. The U.S. Customs Service collects these documents at the port of entry or exit unless the information is filed electronically using the Automated Broker Interface (ABI) on imports or the Automated Export Reporting Program (AERP) on exports.

Census also releases overall trade and transportation statistics that include data elements on the value, commodity, weight, country of origin and destination, and U.S. port used. Many agencies obtain special extractions and tabulations from Census and then perform additional quality assurance reviews and analyses to meet the needs of their own customers. These include: data on North American land trades (by truck, rail, mail and pipeline) released and disseminated by the Bureau of Transportation Statistics as the Transborder Surface Freight Data)[5]; data on U.S. international maritime trade (released to the Maritime Administration and Army Corps of Engineers and disseminated in multiple formats)[6]; and data on U.S. transportation related goods and overall trade data (released to BEA/DOC and disseminated in multiple formats, including balance of payments information). A popular private sector product based on Customs data is PIERS (Port Import Export Reporting Service)[7].

Despite this wealth of information, current U.S. merchandise trade data pose a number of problems when we try to construct origin-to-destination freight movement matrices from them. In particular [4]:

- Current reporting requirements mean that shipping weight is currently only collected for imports, and not for exports.
- In many cases the reported port of entry or exit is not the actual seaport but the port of duty filing, and electronic filing has increased the number of such filings.

---

[5] http://www.bts.gov/transborder/
[6] http://www.marad.dot.gov/Marad_Statistics/index.html
[7] http://www.piers.com/default2.asp

- Data on the domestic origin and destination of international trade is often reported incorrectly. It is not uncommon for origin of movement (OM) series respondents, who may be intermediaries in the goods movement process, to erroneously report either a headquarters location or to specify the location of the US port of exit as the point of an export's origin. The impact is greatest on the allocation of non-manufactured exports, where intermediaries are more common, notably farm products, minerals, and other bulk commodities.
- Due to current reporting requirements merchandise trade statistics do not distinguish goods moved by intermodal combination. Export mode is defined here as simply being the mode used when the U.S. international border is crossed. On the import side, the mode of transportation is defined as the last mode used when the freight arrives at the U.S. port of clearance or entry.

Besides these and a few other issues of content, it is worth noting that anyone unfamiliar with foreign trade data can spend a good deal of time searching the web for specific products. Annual imports and exports by U.S. seaport, foreign port of origin/destination, and commodity class can be obtained. Finding and getting access to them, and associating them with other datasets covering seaport activities can be a challenge. One-site access to at least the most detailed O-D data tables would be beneficial to all users.

3. Data Synthesis Techniques: Some Promising Directions

To the extent that there is any "standard practice" in freight modeling, it involves finding ways to combine data from different sources to support estimation and forecasting of freight movement volumes. References [9-17] identify some recent studies that exemplify the sort of data integration problems we face at the metropolitan/regional [9-12], statewide [13-15] and national/transborder [16, 17] scales. Of note, such efforts include a number of State DOT-based projects that make use of the multi-sourced TRANSEARCH® database; a proprietary product developed by Reebie Associates[8] that offers one approach to what can seem a rather daunting data integration challenge.

Two complimentary lines of attack are suggested for getting the most information out of not only the CFS but also our other freight movement datasets. One employs various mathematical and statistical modeling techniques to merge data from these different sources. The second (see Section 4) involves the use of new surveys/survey design options as well as a set of rapidly evolving non-survey data collection techniques. Given the gaps in CFS coverage highlighted earlier, two important data synthesis activities are discussed below:

- filling gaps in CFS (in- and out-of-scope) commodity flow data;
- creating spatially detailed commodity and vehicle/vessel based freight flow matrices for highways, railways and waterways

[8] http://www.reebie.com/images/transearch.asp This database was also used in the Federal Highway Administration's recent effort to develop a nationwide set of freight movement maps and supporting commodity flow matrices and truck-to-highway assignments [see 16].

3.1 Filling Gaps in CFS O-D Tables

If we think of our data problem as one of filling in a multi-dimensional origin-destination-commodity-mode (ODCM) matrix, a strong candidate for bringing the various elements of such a matrix together is log-linear modeling [18-21]. Mechanically, this approach can be linked to a series of matrix adjustments using the technique of iterative proportional fitting (IPF) which ensures that the values reported or estimated for each cell in the ODCM matrix sum to known or estimated and more aggregate marginal totals. And given forecasts of these aggregate marginal activity totals, the same log-linear models can be used to project the matrix of ODCM cell values into the future.

An example model may help. In this example we assume a single commodity class for brevity and solve for the other three (O-D-M) dimensions. When a region is referred to as generating freight traffic we refer to it as region "i", and when it receives this traffic as region "j". Individual modes are as designated "m". The data product we seek is a fully filled in matrix of freight flows, measured in annual tons moved, $\{F_{ijm}\}$, broken down across each of these three dimensions. For this given commodity we can estimate the following multiplicative model of the tons shipped from region i to region j by mode m, which we would solve in natural log form as:

$$\ln F_{ijm} = \theta + \lambda_i^O + \lambda_j^D + \lambda_m^M + \lambda_{im}^{OM} + \lambda_{jm}^{DM} + \lambda_{ij}^{OD} + \lambda_{ijm}^{ODM} \tag{1}$$

Here the various λ's, often termed the model effects, are a set of model estimated parameters that will return the original cell estimates. For example, the λ_{ij}^{OD} effect returns the impacts of O-D separation on the resulting cell estimate, while λ_i^O represents the size of origin effect. Given a completely filled in flows matrix equation (1) will reproduce the cell estimates exactly. We are interested in how such a model performs with missing data.

In the CFS we will be missing a large number of {ijm} cells: as well as some {ij} and other two–dimensional cells. Setting such cell values initially = 1.0 (log = 0.0) and applying equation (1) we can obtain an estimate of each cell's missing value from the reported cell values. Such estimates are often termed minimum information estimates. Better yet, we can introduce entirely new data into the problem. An appealing feature of this sort of IPF modeling, as described in [19], are the many possibilities for treating missing data elements. In particular, we can combine data from the CFS matrix with data from other sources, such as the railcar Waybills (suitably modified to match CFS regions and commodity classes). Here this Waybills data can be used as a second estimate or "data model" of the rail flows in each commodity class. We can do this in a number of ways. We can replace CFS-based missing cell data with waybill estimated values and then use IPF to bring the full matrix (in all four dimensions) back into compliance with the original CFS flow margins. We can also treat the railcar waybill flows as though they were a separate dimension or set of commodity specific tables in the rail portion of the CFS flows matrix, and fill in the missing valued cells using a combined CFS and waybills-inclusive log-linear model.

We can carry out the same operations on those parts of the commodity flow matrix involved with water and air freight transportation, for example using Corps of Engineers and Office of Airline Information data respectively, in place of the railcar Waybill information as the second "model" of these flows. Finally, we can carry out a similar operation for truck shipments: but in this case we will have to substitute an actual model of flows in place of a second data set. An additional possibility here is to incorporate a set of travel distance intervals as yet another dimension into the log-linear modeling solution, broken down by commodity class. The CFS reports tons moved in the mileage ranges less than 50, 50-99, 100-249, 250-499, 500-749, 750-999, 1000-1499, 1500-2000, and over 2000 miles.

3.2 Constructing Spatially Detailed O-D Matrices

Construction of Spatially Detailed Truck Trip Matrices is perhaps the most pressing data need at the present time for both statewide and metropolitan area planning. Without the equivalent of a spatially explicit railcar waybill or inland waterways vessel manifest, truck trip data falls to the CFS, the VIUS, route specific truck traffic counts, and any local trucking survey data that may (very occasionally) become available. CFS data is best suited to statewide analyses involving long-haul freight activity. It can however offer some useful regional control totals, as well as value-to-weight statistics, for use in metropolitan area studies. The following discussion assumes a statewide or similar (e.g. multi-state corridor) type of planning application.

One way to create O-D flows at a level of spatial detail finer than CFS regions is to combine CFS data with county-based economic activity data. This involves allocating CFS-based freight activity across counties based on the volume (earnings, employment levels) of freight-generating industrial activity reported by Economic Census. One way to accomplish this is to pass this economic activity data through a regional input-output model that translated dollars of industrial activity into dollars of commodities produced and consumed [22 - 25]. Recent Census releases of annual employment and payroll data at the 5-digit zip-code now allow modelers to consider using different within-CFS regional aggregations of this data that may in some cases (some commodities) be better suited to the problem of generating spatially explicit freight flows than counties, especially if the next step is to turn these flows into trucks and assign them to the highway network.

Input-Output (I-O) software, including well supported and commercially available codes, is now generally available, allowing analysts to associate the inputs and outputs of different commodities with specific industries. Where data on the annual production of a specific commodity class already exists (e.g. bushels of grain produced in a county) the principal use of I-O modeling is in the consumption, and therefore destination, end of an O-D movement. What is required is a commodity-to-industry conversion table that associates the amount of each commodity required to produce a unit of that industry's output. We can represent this as follows. Let u(g,n) refer to the sales of commodity g to industry n. Given this data for all n= 1, 2, … N industries in a region/county/zip code area, we can compute Q(g), the total quantity of commodity g consumed as:

$$Q(g) = u(g,1) + \ldots + u(g,N) + e(g) \qquad (2)$$

In practice further adjustments to the data are usually needed. Getting county-level or zip code area tonnages from this approach then means using CFS provided ton-per-dollar shipped statistics. Getting the number of truck trips from these tonnage values also requires suitable data on average truck payloads, statistics that need to include the percentage of empty as well as fully or partially loaded truck trips (e.g. backhauls) involved: see [15] and [16] for example applications, both using VIUS truck weight data.

Given such a spatially disaggregated set of commodity specific productions (Os) and consumption totals (Ds), the next step is to link the two to create a set of O-D flows. One way to do this is to calibrate a set of commodity specific spatial interaction models, subject to CFS region-to-region control totals. Doing so, however, requires additional information, such as the average distance shipped (per ton) or a distribution of trips over different distance intervals. Again, CFS can provide both of these types of statistic, but for limited commodity or regional breakdowns.

Given a set of truck flows there is then the issue of validating the resulting estimates. About the only truck activity data available for this purpose in most regions is average annual daily truck traffic (AADTT) counts, including the counts States report to the Federal Highway Administration as part of their HPMS submissions.[9] If these counts can be grouped into appropriate O-D specific traffic corridors then some level of aggregate truck trip comparisons might be attempted. However, for this approach to be meaningful requires both a reliable set of representative truck counting sites and counters with the ability to identify trucks of different sizes and axle configurations. It also requires a reliable means of assigning commodities to truckloads. A more careful look at how truck count data might be used is in order. A number of mathematical programming models have been developed that allow modelers to combine O-D and link count data to estimated truck movement matrices [26 - 28]. Applications to date have focused largely on the intra-urban scale. For example, see [27], which describes a flexible approach that lets the analyst place greater reliance on either the O-D or count data, as warranted.

Construction of Rail, Water and Intermodal Trip Matrices usually occurs as part of long-haul freight studies. Here the CFS can play an important role in support of the modally more comprehensive (in commodity coverage terms) Waybills and Waterborne Commerce datasets. Among other things the creation of rail or waterway inter-regional O-Ds means dealing with the issue of truck drays to and from these line-haul modes. Here the CFS offers some annual region-to-region control totals. It might also be used in the future to provide distributions of estimated truck-rail, truck-water and also truck-airport drayage distances. Data on such draying activity is of considerable importance to States where a significant volume of intermodal freight crosses their borders. Truck draying costs, and hence distances, can also be important in determining the geographic size of the shipper "market area" served by water and rail modes, and by specific airports. These drayage distances are currently estimated for the CFS by Oak Ridge National Laboratory. How well they reproduce actual drays needs to be

[9] Highway Performance Monitoring System. http://www.fhwa.dot.gov/policy/ohpi/hpms/

established. The key unknown here is the true location of the rail or river loading/unloading dock for each shipment. Perhaps one or more of the data collection approaches discussed below can help with this question.

4. New Approaches to Data Collection: Some Promising Directions

With funds for freight data collection limited at all levels of government we need to take full advantage of any opportunities that come along to fill in the freight movement picture. The following are three areas worth exploring:

4.1 Linking CFS, VIUS and Other Trucking Data

As identified in [4] the principal options for sampling truck trip activity patterns, besides a shipper survey such as the CFS, are a) vehicle-based sampling, b) vehicle/driver intercept surveys and c) vehicle tracking surveys. Each approach has its pros and cons. The costs of options b) and c) render them unrealistic as nationwide sampling options. Vehicle identification number-based sampling is currently used by the VIUS, and here a number of interesting options may be worth researching, such as:

- the collection of truck activity diaries for a sub-sampling of vehicles covered by the VIUS, focusing on vehicle routing, backhauling, repositioning and operating speed aspects of the freight pickup and delivery operation.
- a tie-in between a more continuous truck-trip dairy based survey and the currently quinquennial VIUS, making use of the latter's sample frame.[10] Diaries might be collected on an annual, rotational basis, possibly with different types of vehicles or commodities selected each year for diary completion. The characteristics of the vehicles reported in these travel diaries might then be tied to operating characteristics in the larger, 5-year vehicle sample.
- a more radical approach might include a redesigned CFS, such as an alternating bi-annual shipper (CFS) and motor vehicle (VIUS) survey program, with the latter including some form of truck trip diary sampling.

It is also worth exploring how we might tie local truck/commercial vehicle activity surveys, as occasionally collected by MPOs, to such developments in data collection.

4.2 The International Trade Data System: How (and When) Can We Use It?

International freight data collection in the United States is currently undergoing a major change. In the near future international trade data will flow to the US DOT and some 104 other federal agencies (at last count) via the International Trade Data System (ITDS)[11]. ITDS is a federal government information technology initiative (Initiative IT06) to coordinate, standardize, and ultimately simplify our Federal border clearance and other international trade

[10] For the 2002 VIUS the Census Bureau purchased a sample frame generated by R.L. Polk & Co.
[11] http://www.itds.treas.gov/sitemap.html

and transportation processes. It will enhance and replace the Automated Commercial System (ACS) currently used by the Bureau of Customs and Border Protection within the Department of Homeland Security. Traders will submit standard electronic data for imports or exports only once, with the ITDS system serving as a "single window" system through which trade transactions data can flow between private traders and over 100 federal agencies involved in international trade.

The ITDS is a timely response to the federal government's need for much greater visibility of incoming foreign cargos in this time of heightened concern over terrorist actions. To this end each federal agency submits a list of data elements it deems key to its operations. The US DOT is developing five different portals into the ITDS: one each for MARAD, FMCSA, NHTSA, FAA, and a BTS supported and maintained portal that will supply freight movement data to other modal administrations within the US DOT (FHWA, FRA). It remains to be seen how and when this data becomes available for general use; and raises the question of how such data might be used to study complete end-to-end movements of freight through international supply chains.

4.3 Freight Informatics: Cargo Tracking, Supply Chains and Electronic Manifests

Freight Informatics. The growing presence of many different kinds of real time information gathering technology means that future traffic data collection is going to make greater use of non-survey based approaches. Hopefully these methods will prove less expensive than traditional survey methods in the not too distant future. Existing technologies include the following [4]:

- *Active Roadway Sensors*: Fiber Optic Sensors, Inductive Loop Detectors, Magnetic Sensors, Piezoelectric Sensors, Pneumatic Road Tube, Weigh-in-Motion (WIM) Sensors;
- *Passive Roadway Sensors*: Infrared Sensors, Microwave Radar, Passive Acoustic Array Sensors, Ultrasonic Sensors, Video Image Sensors
- *On-Board Sensors:* Bar-Code Scanners, Microchip-Based Smart Cards, Radio Frequency Identification Devices (RFID) and Remote Intelligent Communication (RIC), Smart Active Labels, Satellite/GPS-Based Vehicle Tracking.
- *Wide-Area Sensors*: IKONOS satellite imagery, Light Detection and Ranging (LIDAR), small plane, helicopter and uninhabited autonomous vehicles/micro aerial vehicles (UAVs/MAVs)

An RFID (Radio Frequency Identification Device) system typically consists of a tag or label containing data storage, an antenna to communicate with the tag, and a controller to manage the communication between the antenna and the computer. An RFID tag can be embedded in a package or placed on a person. Combined with remote intelligent communication (RIC) technology, RF-based wireless reporting can be used to track the location, condition, and content of goods at every stage in a product‘s supply chain, and do it in near real time. This includes the emerging technology of Smart Active Labels (SALs) which use RFID tags containing an internally powered microchip linked to an antenna for wireless reception and transmission purposes. A read/write mode, suitably powered RFID can be used as a dynamic, electronic cargo manifest. The potential for increased cargo security alone is going to bring this sort of “smart tag” technology into the mainstream for freight and inventory management. A now widely used tracking technology is GPS. The commercial component of the Global Positioning System (GPS) is a worldwide radio-navigation system formed by linking together 24 orbiting satellites and their network of ground stations. Vehicle or cargo tracking down to a few meters is already possible, with further spatial refinements (down to centimeters) under development. Useful tracking of vehicles might base sampling on high volume highway corridors or high volume freight gateways, possibly on an annual, rotational basis. The use of micro aerial vehicles (MAVs) seems likely to bring down the price of surveillance at major traffic intersections or along major traffic corridors. Here the potential for freight flow analysis would be in the combination of traffic count data from these aerial devices with O-D survey data, possibly as joint inputs to the “link O-D models” mentioned in section 3.2 above.

The potential for largely automated freight data collection seems obvious, given enough time and resources currently being used to develop informatics technologies. Less obvious, and in need of study, is the use to which this information can be put by public agencies. We also need to ask what other information technology is out there and what other uses we can find for those listed above.

Tracking Freight Supply Chains. Hopefully the ITDS program will eliminate the current weaknesses in our import/export data. If it can do so, then we might also ask whether a similar effort might not be used to collect the physical origin-to-destination movements of domestic cargos. A logical next step would then be to combine both of these domestic and international datasets. Perhaps the ultimate expression of this idea is the creation of a universal electronic manifest (UEM). Such a manifest would provide essential information for shippers and carriers to manage inventories and logistics, as well as meet the documentation needs of domestic and international trade, hazardous cargo movement, and the growing security needs of domestic transportation. It would modernize existing paper-based waybills and allow cargo tracking across all modes.

Unless or until the ITDS or another UEM-based system evolves, we will have to rely heavily on data integration tools such as those discussed above to piece together this goods movement puzzle. Recent U.S. experiments with an electronic supply chain manifest suggest that information technology applications that benefit private freight movement agents will advance quickly. A goal for the Federal Motor Carrier Safety Administration’s (FMCSA) Commercial Vehicle Information Systems and Networks (CVISN) program [29] is the integration and automatic processing of multi-sourced data on the carrier, vehicle, driver and cargo (including oversize and hazard designations) associated with domestic or international commercial truck trips. A CVISN link to the ITDS/Automated Commercial Environment (ACE) has also been discussed. This sort of data integration activity across federally supported programs needs further exploration. What other useful connections can we find if we look at the whole breadth of the government’s IT-based information gathering activities?

Similar experiments with different freight IT applications are ongoing in Europe [30]. How might these different technologies be used to develop aggregate statistics for use by participating public agencies, notably in the estimation of hourly, daily and (through aggregation) seasonal O-D freight flow volumes? There is also considerable potential, some of it already being tapped, for monitoring and measuring the travel speeds and en route delays associated with location specific truck, rail and waterborne commerce movements.

Such possibilities also suggest an alternative approach to survey design: the use of "supply chain surveys", involving a mixed sample drawn from a mix of establishment types, i.e. from shippers, carriers, distributors, terminal operators, receivers and also freight forwarders. More than one U.S. experiment in the tracking of complete O-to-D supply chains is currently underway, again making use of recent developments in electronic reporting. An experiment with this type of data collection was also recently tried in Europe, based on more traditional data collection methods involving both face to face and telephone interviews with the different supply chain participants [31]. This sort of survey offers a greater understanding of the full logistics costs involved in moving freight from source to destination. However, it also requires a potentially complex, multi-stage sample design. Both a benefit (and an added complication) of such a survey is the collection of data on who is responsible for moving the freight at each stage, and what institutional arrangements (e.g. between carrier and shipper, shipper and broker, shipper and receiver) have been made to facilitate this. The European experience identifies a number of problems with achieving good response rates, and with higher costs per successful response than with traditional shipper surveys [31]. In particular, the probability of getting a complete description of complex, multi-actor supply chains proved to be low. How such an approach might stand up under different sampling designs and under mandatory reporting requirements is currently unclear.[12]

5. Summary

The patchwork quilt that is today's freight data coverage requires a good deal of effort to create even a base case set of commodity flows for use in regional planning studies. Spatial disaggregation of flows requires that the CFS be combined with other datasets. Combined commodity and spatial disaggregation within the CFS is especially limited in this context. A number of quantitative methods can be used to help us get the most out of both the CFS and other federally supported freight movement datasets. New ways of collecting freight data, and especially data on truck movements, offer the promise of not only greater coverage but also greater understanding of how freight really moves through the transportation system. Directions for future research and development identified in the paper include:

- the use of iterative proportional fitting and log-linear modeling techniques, including spatial interaction modeling, to produce maximum likelihood estimates of empty cell values in commodity and mode specific freight movement matrices
- the use of economic activity data and spatial input-output models to disaggregate truck trip commodity flow matrices
- the tracking of complete source-to-final market product supply chains, following the ownership as well as the physical and geographic aspects of en route cargo transfers

[12] Even if the average response rate is 75% at each stage in the supply chain (a little higher than the rate experienced by the CFS) a three-interview chain would yield only a 42% successful completion rate, assuming independence of responses at each stage

- the measurement of truck draying distances associated with intermodal shipments
- the use of both CFS and foreign trade data to develop spatially detailed estimates of the true origins, destinations, and modes used when transporting imported and exported goods, converted to both tons and dollar valued trades
- the development of functional linkages between the CFS, VIUS (possibly expanded to include truck trip diaries) and perhaps also local area truck trip activity surveys
- the gradual incorporation of vehicle counting and vehicle tracking information into the estimation and validation of truck-based commodity flows
- the rationalization of foreign trade data statistics in support of mode and O-D tracking

The need for better freight movement data is considerable, and the costs of using poor quality data to plan future investments are likely to rise along with the pressure such movements place on our transportation infrastructures. Perhaps one final research task ought to be an objective assessment of such costs versus the benefits of providing better data to public agency planners and decision-makers.

References

1. TRB (2003) A *Concept for a National Freight Data Program*. Special Report 276. Transportation Research Board, Washington, DC 20001. http://gulliver.trb.org/news/blurb_detail.asp?id=1730
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. (http://www.dot.state.ny.us/tss/conference/synthesis/pfd
3. Southworth, F. (1999) *The National Intermodal Transportation Data Base: Passenger and Goods Movement Components*. Report prepared for the Bureau of Transportation Statistics, U.S. Department of Transportation, Washington, D.C. 20590.
4. 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.
5. 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.
6. 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
7. Capelle, R.B. et al (2000) *Freight USA*. Highlights from the 1997 Commodity Flow Survey and Other Sources. Report prepared by Oak Ridge National Laboratory for the Bureau of Transportation Statistics U.S. Department of Transportation, Washington D.C 20590.
8. U.S. Census Bureau (2002) 2002 United States Economic Census. *Vehicle Inventory and Use Survey*. EC02TV-US. http://www.census.gov/svsd/www/02vehinv.html
9. Gordon, P. and Pan, Q. (2001) Assembling and processing freight shipment data: developing a GIS-based origin destination matrix for southern freight flows. University

of Southern California, Los Angeles, CA.
http://www.metrans.org/Research/Final_Report/99-25_Final.htm
10. Clarke, M. and Cheng, L. (2005) Initial development of an innovative commodity flow and truck model for the Greater Los Angeles Basin. International Chinese Transportation Professionals Association Annual Conference, Washington DC. http://www.ictpaweb.org/publication/DC/fu-2005conpaper/1-1-4-cargotechnicalpaper.doc
11. VPRA Associates and Cambridge Systematics Inc. (2002) Goods Movement Truck Count Study. Report prepared for the Southern California Association of Governments, Los Angeles, CA 90017.
12. Liu, N.L. and Vilain, P. (2004) Estimating commodity inflows to a substate region using input-output data: commodity flow survey accuracy tests. *Journal of Transportation and Statistics 7.1*: 23-37.
13. Brogan, J.J., Brich, S.C. and Demetsky, M.J.(2001) Application of a Statewide Intermodal Freight Planning Methodology. Virginia Transportation Research Council Report VTRC 02-R5.Charlottesville, VA. http://www.virginiadot.org/vtrc/main/online_reports/pdf/02-r5.pdf
14. Parsons-Brinckerhoff-QD, Inc. (2002) Developing A Framework To Assess Goods Movement in New Jersey. Technical Appendix. Report prepared for the New Jersey Department of Transportation, Trenton, NJ 08625. http://www.state.nj.us/transportation/works/freight/documents/GMFtechApp.pdf
15. Beagan, D. and Grezenback, L. (2002) *Freight Impacts on Ohio's Roadways.* Report FHWA/OH-2002/026 Ohio Department of Transportation. Columbus, OH. http://www.dot.state.oh.us/planning/Freight/freight_impacts_report.htm
16. Battelle Team, Reebie Associates, Wilbur Smith Associates and Global Insight (2002) *Freight Analysis Framework: Derivation of the FAF Database and Forecast.* Report prepared for the Federal Highway Administration, Washington, D.C. 20590. http://ops.fhwa.dot.gov/freight/documents/WebMeth/index.htm#1998
17. BTS. (2003) *Measurement Of Ton-Miles And Value-Miles Of International Trade Traffic Carried By Highways For Each State. International Trade Traffic Study* (TEA-21, Section 5115). Bureau of Transportation Statistics, US Department of Transportation, Washington, DC 20590.
18. Willekens, F. J. (1983) Log-linear modeling of spatial interaction. *Papers of the Regional Science Association, 52*: 87-205.
19. 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).
20. Agresti, A. (1990) *Categorical Data Analysis*. Wiley, New York
21. Wrigley, N. (1985) *Categorical Data Analysis for Geographers and Environmental Scientists*. Blackburn Press, England.
22. Sorratini, A.J. and Smith, R.L. Jr. (2000) Development of a statewide truck trip forecasting model based on commodity flows and input-output coefficients. Transportation Research Record 1707: 49-55
23. Liu, N.L. and Vilain, P. (2004) Estimating commodity inflows to a substate region using input-output data: commodity flow survey accuracy tests. *Journal of Transportation and Statistics 7.1*: 23-37.

24. Jackson, R. W et al (2004) A method for constructing commodity by industry flow matrices. Research Paper 2004-5. Regional Research Institute, West Virginia University, Morgantown, WV. http://www.rri.wvu.edu/pdffiles/jacksonetal2004-5wp.pdf
25. Canning, P. and Wang, Z. (2004) A flexible modeling framework to estimate interregional trade patterns and input-output accounts. *Journal of Regional Science, 45(3).*
26. Van Zuylen, J.H. and Willumsen, L.G. (1980) The most likely trip matrix estimated from traffic counts. *Transportation Research 14B*: 281-293.
27. List , G.F. and Turnquist, M.A. (1994) Estimating truck travel patterns in urban areas. *Transportation Research Record 1430*: 1-9.
28. Sherali, H.D., Sivanandan, R. and Hobeika, A.G. (1994) A linear programming approach for synthesizing origin-destination trip tables from link traffic volumes. *Transportation Research 28B*: 213-223.
29. FMCSA (2002) Commercial Vehicle Information Systems and Networks (CVISN). Recommendations for Primary Identifiers. Baseline Version V1.0. White Paper. Federal Motor Carrier Safety Administration. Washington DC 20590.
30. EU (2000) Sustainable Urban and Regional Freight Flows (SURFF). EU/DG XIII Transport Telematics Applications Program. TR 1053. European Union. http://www.euroweb.net/surff/about.htm
31. Rizet, C., Guilbauly, M., van Merijeren, J.C., Bjister, E. and Houee, M. (2003) *Tracking Along The Transport Chain Via The Shippers Survey*. International Conference on Transport Survey Quality and Innovation. http://www.its.usyd.edu.au/