Integrating heterogeneous data sources for better freight flow analysis and planning

José Luis Ambite, Genevieve Giuliano, Peter Gordon, Qisheng Pan, Sandipan Bhattacharjee
University of Southern California, Los Angeles, CA 90089
ambite@isi.edu, {giuliano,gordon,qpan,sandipab}@usc.edu

Abstract

We present ongoing work on developing an automated integration system for freight flow analysis and planning. To overcome the limitations of current estimation methods for commodity flows, we use reliable secondary sources, such as employment data, and derive estimates in a principled way by means of a computational workflow. When available, we extract the data automatically from online sources, so that the system maintains the estimations continuously updated. This project will allow planners and policymakers to make more informed decisions by accessing the most recent data and enhance the ability to explore different scenarios.

1 Introduction

Economic restructuring and globalization have vastly increased the volume of commodity flows by all transport modes. In the US, intercity ton-miles have increased approximately with GNP, but truck and air transport have increased faster than other modes. In 1994, for example,, of the total bill of $420 billion the US spent on transportation, trucks carried 79%, i.e. $331 billion (USDOT, 1998). Total US ton-miles of freight increased from 2989 billion in 1980 to 3710 billion in 1998. Over the same period intercity truck ton-miles almost doubled (from 555 to 1027 billion), and domestic air ton-miles increased from 4.5 to 13.8 billion (USDOT 2000).

Increased freight flows have had significant impacts on metropolitan areas. Traffic at major freight generators (ports, airports, rail yards, warehouse/distribution nodes) has greatly increased, adding to congestion and impacting surrounding neighborhoods. Increased train traffic interrupts road traffic. Increased truck traffic has accelerated deterioration of highways and often conflicted with demands for passenger commuter service. In addition, rapid changes in economic linkages are leading to ever changing flow patterns and the spatial restructuring of metropolitan areas (Graham and Marvin, 1996; Gordon, Richardson and Yu, 1998; Giuliano 1998).

As freight flows and their impacts increase, transportation planners, managers and operators have a greater interest in developing better methods for tracking and monitoring commodity flows, and for analyzing these flows as they impact transportation nodes and networks. Yet, current freight flow estimation and analysis methods have several problems, some related to data and some related to the estimation methods themselves. This paper presents a new approach for commodity flow estimation in metropolitan areas. We use a regional input/output model and combine it with available import/export commodity flow data to estimate detailed commodity flow matrices. Additional computations allocate flows to modes and ultimately assign flows to the transportation network. Our approach utilizes plausibly reliable data sources, manages to integrate heterogeneous data, and provides a means for validating and calibrating network flow estimates. Use of data integration and automation techniques should make possible continuously updated and detailed freight flow estimates.

The remainder of this paper is organized as follows. Section 2 describes current freight flow methods and their problems. Section 3 presents the conceptual framework for our model. We discuss the motivations for our "bottom-up" approach and describe the model as developed to date. Section 4 presents our plans for constructing and testing a continuously updatable freight flow estimation model.

2 Current Methods for Freight Flow Estimation

There is an extensive literature on urban transportation network modeling, and the state-of-practice is well advanced (Wilson 1970; List and Turnquist 1994; Willumsen 1978, 1984). However, such models (except List and Turnquist 1994), do not explicitly treat freight flows. The usual method of modeling truck flows in metropolitan area analysis, for example, is to use rule-of-thumb fixed factors based on passenger vehicle flows and observed truck counts at a small number of locations on the highway network. Rail freight flows are not usually modeled at the metropolitan level.[1] This simple approach was adequate when trucks accounted for only a small percentage of urban traffic, and when regional planners and policymakers were relatively unconcerned about intra-metropolitan freight traffic.

The situation has now changed, particularly in large metro areas like Los Angeles, home of the largest container ports (trade in 2000 was $200 Billion) as well as the second largest air freight airport in the US. Globalization, restructuring of goods supply chains, and changes in warehousing practices have resulted in large overall increases in freight traffic, and extremely large increases associated with ports and airports. Increased freight traffic has, in turn, led to a number of problems: highway congestion, traffic accidents, roadway deterioration, air pollution, noise, risk due to hazardous materials transport, etc. In such cases freight modeling methodologies have to be more exacting. The requirements for such a model are: (a) have solid behavioral foundations, (b) be multimodal, (c) should be able to analyze interactions between passenger and freight trips and (d) able to take feedback from policy changes (Hedges, 1971).

Urban researchers have a limited understanding of freight flows, both because passenger transport has been the dominant concern and because of the dearth of detailed freight flow data. However, there is growing interest in trying to estimate and understand commodity flows for many reasons, e.g. costs and impacts of commodity flows on regions and local areas; relationships between supply chains, flows and firm location behavior; costs and benefits of international trade. Yet, detailed origin-destination studies are expensive, often resisted by shippers (many of whom treat the data as confidential) and, therefore, difficult to update. There are a wide variety of approaches to reach freight modeling. They range from simplistic techniques (like, estimating regression models between freight and passenger flows; assuming that freight follows same patterns as passenger flows), adaptations of passenger models to freight transportation analysis, and in very few cases specific freight models have been developed.

2.1 Data Problems

Current methods of freight flow estimation have many disadvantages mostly due to lack of data that is appropriate, accessible, and reliable. Ideally, one would like to have accurate data on commodity flows by industry sector, mode, origin and destination at a geographic scale sufficiently fine to identify flows on specific routes or at specific locations. Since a large part of flows within a region either originate or are destined to locations outside the region, the regional import/export component is critical. Such a comprehensive data source does not exist, leaving analysts with two choices: develop an estimation method based on available data, or collect the necessary data directly from freight transporters. Any survey approach to collect data from trucking companies, railroads, air transport firms, etc. would be prohibitively costly, even if private firms were willing to provide their proprietary information. Moreover, freight flows vary over time, and hence would require repeated surveys.

Reliance on conventional secondary data sources has its own problems. Metro level analysis requires fine geography; most existing data is at regional scale or higher. With respect to commodities, there are various classification systems, units (dollars, tons), varying levels of aggregation, more information on import/export flows, little information on intra-regional shipments; more data on port, air import/export, little data on truck, rail imports/exports.

[1] Commodity flows are typically modeled at the inter-regional or inter-state level.

There are also problems associated with how to account for empty trucks, warehouse/secondary processing activities, intermodal exchanges within any region and how to account for data collected at different times, time intervals. While we cannot solve all the problems, Section 3 summarizes how we developed ways to circumvent many of them.

2.2 Standard Approaches and Methodological Problems

The Quick Response Freight Manual released by the US Department of Transportation in 1996 provided simple techniques and transferable parameters for developing commercial vehicle trip tables (USDOT 1996). Truck trip generation rates are estimated from the number of jobs in the employment sectors that are associated with commodity shipments. The default rates provided by the manual were taken from a survey in Phoenix, Arizona. After calculating truck trips from employment data, it is straightforward to construct a truck trip table and assign the trips by following the conventional UTPS (Urban Transportation Planning System) four-step models. This method is simple to implement. However, it is a nagging problem that the default parameters like truck trip generation rates are not easily transferable between different regions.

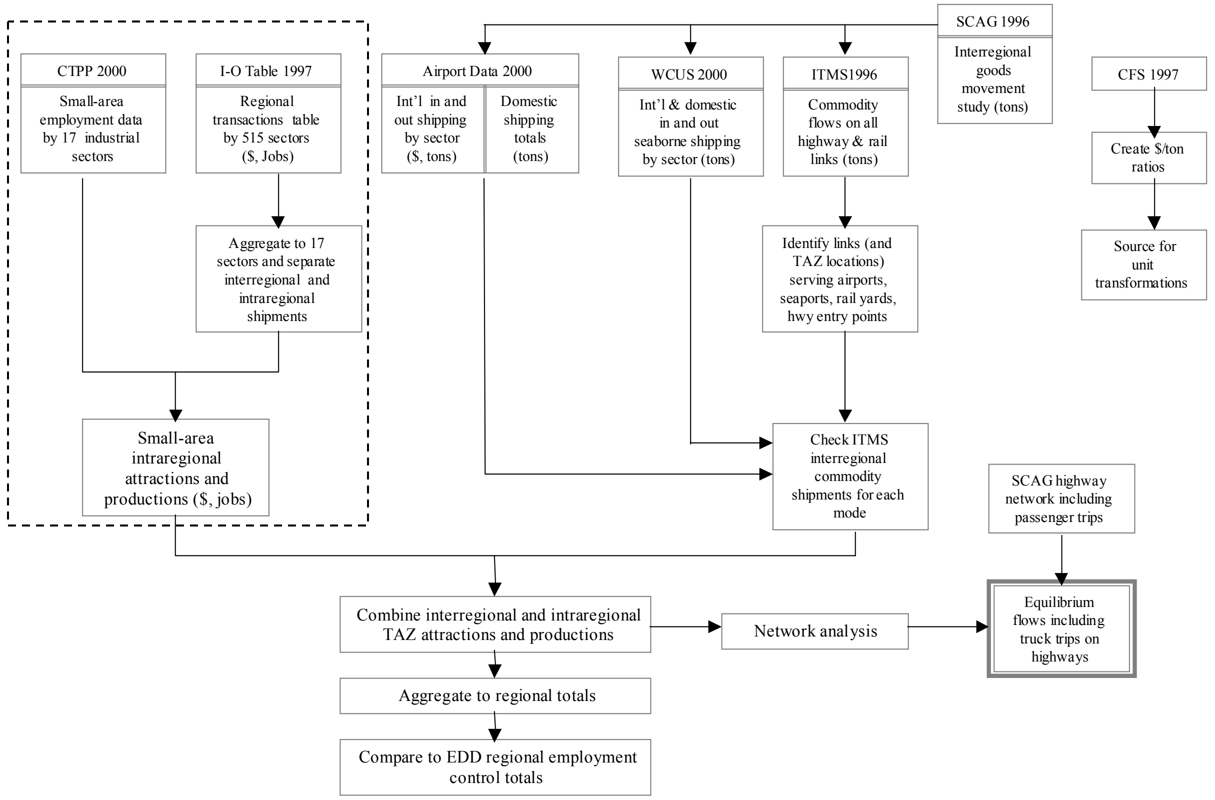
In the Los Angeles area, the Southern California Association of Governments (SCAG) developed a Heavy Duty Truck (HDT) Model in 1999 to forecast HDT travel patterns, traffic volumes as well as Vehicle Miles Traveled (VMT) for the entire SCAG region. The forecasts for HDT activities were based on truck HDT trip generation rates developed through surveys, regional economic data and commodity flow data, and the activity at special generators, such as airports, seaports and intermodal transfer facilities. After trip generation, the HDT trips were distributed using gravity models with the friction factors estimated from truck trip diary survey data. At the end, the HDT trips are assigned to regional highway networks and VMTs are estimated for emissions analyses (SCAG 1999). As with the Quick Response Freight Manual, SCAG's HDT model also used employment data to estimate internal truck trip generation rates on the basis of shipper-receiver survey data. However, the survey sample size was small and the survey was conducted over a short period of time due to limited funds.

The California Department of Transportation (Caltrans) has released three versions of the Intermodal Transportation Management System (ITMS) for statewide transportation planning since 1996. ITMS estimated freight movement by different modes based on data from Interstate Commerce Commission's 10% waybill sample, Reebie Associate's Transearch database, Dri/McGraw Hill's U.S. economy forecast, and Port Import Export Reporting Service (PIERS). The ITMS traffic analysis zones are based on existing zip code areas. Because the Transearch database provides commodity flow information for different modes, ITMS has been working on converting commodity flows into truck trips and assigning the trips to the state-wide transportation network. The freight model part of ITMS is still under development

2.3 Conceptual Framework for Integrated Model

Problems with these approaches as well as the various data limitations motivate our approach to estimating detailed freight flows. For purposes of discussion, there are two possible ways to proceed: Bottom-up or top-down. The bottom-up approach derives the flows from the underlying economic activities that generate demand and supply. A top-down approach uses available flow data to estimate total flows. Problems with top-down include a lack of detailed network flow data, (truck, rail) and no easy way to validate commercially available commodity flow data (REEBIE). We use a bottom-up approach that utilizes available small-area employment data, a relatively reliable measure collected at regular intervals and often available at a reasonably disaggregate level.

This work builds on a suggested approach to the problem by Gordon and Pan (2001; summarized in Figure 1). They presented a prototype case study of the Los Angeles metropolitan area, a region that includes the twin ports of Los Angeles and Long Beach that together top the U.S. in terms of container shipments, as well as the Los Angeles International airport among others.



SCAG 1996
Interregional goods movement study (tons)
CTPP 2000
Small-area employment data by 17 industrial sectors
I-O Table 1997
Regional transactions table by 515 sectors ($, Jobs)
Airport Data 2000
Int'l in and out shipping by sector ($, tons)
Domestic shipping totals (tons)
WCUS 2000
Int'l & domestic in and out seaborne shipping by sector (tons)
ITMS1996
Commodity flows on all highway & rail links (tons)
CFS 1997
Create $/ton ratios
Source for unit transformations
Aggregate to 17 sectors and separate interregional and intraregional shipments
Identify links (and TAZ locations) serving airports, seaports, rail yards, hwy entry points
Small-area intraregional attractions and productions ($, jobs)
Check ITMS interregional commodity shipments for each mode
SCAG highway network including passenger trips
Combine interregional and intraregional TAZ attractions and productions
Network analysis
Equilibrium flows including truck trips on highways
Aggregate to regional totals
Compare to EDD regional employment control totals

Figure 1. Summary of Freight Data Collection and Processing.
(The dashed area is shown in detail in Figure 2)

The approach begins with an input-output model of the local economy and divides transactions into two commodity flows types, intra- and inter-regional; a transactions table less interregional trade is created. The modified transactions table is used to create two coefficients matrices: Traditional Leontief coefficients as well as a matrix of their mirror opposite, the sales-based coefficients. Combining these with the available small-area jobs data, Gordon and Pan calculated the commodity values associated with each of the region's 1527 Traffic Analysis Zone's intra-regional supply and with each zone's intra-regional demand. Summing, they get shipments produced and shipments attracted by and from each zone by sector for purposes of intra-regional trade.

Because various data sources had to be combined, it was often necessary to convert units -- from tons to dollars to jobs to ton-miles to container units to trucks to passenger-car-equivalents, etc. One indispensable source for many such conversions was the U.S. Commodity Flow Survey (1997), which provided dollars per ton data for a large number of industrial sectors and for different modes; these data are available every five years at the state level.

This leaves the estimation of shipments to and from selected zones for purposes of interregional trade. This calculation only concerns a limited number of zones: the two major seaports (Long Beach and Los Angeles); the five major airports involved in freight shipping; three major railyards and six major highway entry-exit points. The zones thereby identified are the ones involved in shipping beyond their intra-regional role because these select zones have jobs associated with both types of trade. The estimated inter-regional trade data is also useful to update and check the ITMS data, which theoretically include all of the flows going in and out of these zones on all of the relevant major highway links.

A variety of data sources were used for the estimation of interregional trade. LAX and (to a lesser extent) Ontario airport have substantially detailed data on international shipments. National shipments data are not as easily available. Statewide proportions from the Commodity Flow Survey may have to be used. In the Gordon-Pan application, some rule-of-thumb proportions from a SCAG study were utilized.

Likewise, data for waterborne trade are available from Waterborne Commerce of the United States (WCUS). The 1996 seaborne commerce data for the Long Beach and Los Angeles seaports were downloaded from the WCUS web site (http://www.wrsc.usace.army.mil/ndc/wcsc.htm). The WCUS data are tabulated by Standard Industrial Trade Classification (SITC) categories. The first necessary step was to aggregate the classes of SITC commodities to the freight sectors used in Gordon-Pan study.

SCAG's 1996 Interregional Goods Movement Study is a special resource available to our research because of a major effort by SCAG and its consultants through the early 1990s. The SCAG study provided the total tonnage of shipments in and out of the region by mode, and the ratios of goods originated in or destined to the SCAG region.

The completed origin-destination matrix denoted a sum of 6.21 million metropolitan area jobs, a sum reasonably close to the actual 1990 figure of 6.61 million jobs. Reliable County-level employment aggregates are available monthly from California's Economic Development Department. More useful tests would have to take place at the individual link level. Actual freight flows for selected links are known from various screenline surveys. Yet, before any such testing could occur, the various freight flows had to be assigned to the region's highway network. The challenge was to add these new flows (in passenger-car equivalents) to SCAG's estimated passenger flows on all links.

Traffic assignment models the trip-maker's choice of path between all available zonal pairs. Equilibrium-based travel demand models are usually adopted for the purpose of traffic assignment. For a congested network condition, strict network assignment models are appropriate to predict the equilibrium flows. Based on the theory of User-Optimal-Strict (UO-S) On Network Assignment (NA), Sheffi (1985) provided a traffic assignment model that assumes perfect rationality among travelers, no temporal fluctuations and no modal or link interactions. Sheffi's method was implemented to assign the passenger and truck trip volumes to the highway network of the SCAG region. Applying this algorithm to the minimization of the UO-S model requires a solution of all feasible values to be generated at each iteration step. When the results become convergent, the total travel time on the network is minimized, assigning all trips to the shortest travel-time path of the origin destination pairs.

In summary, Gordon and Pan have developed a methodology for estimating the detailed commodity flows based on secondary sources. They use reliable secondary sources, as available, and derived the estimates in a more plausible way than the competing approaches discussed in Section 2. Nevertheless their method is a mixture of manual and computer-based calculations. Since most of the data sources exist in electronic form, often as databases or web sources, we describe in the next section how we plan to fully automate the estimation process.

3 Automated Integration and Estimation of Freight Flows

Our current work improves Gordon and Pan's approach in several ways. First, we seek better and more compatible data sources as they become available. Second, we automate the data extraction from online sources. For example, if the data are available at the web site of an agency, we can extract the data automatically and maintain the data continuously updated. Third, we automate the processing steps that derive the desired estimations. These calculations involve database manipulation operations, such as selection, projection, and join, aggregation operations, spatial operations, such as changing data from a zoning system to another, and complex algorithms, such as the imputation of trips to specific links in a highway network. Finally, we plan to perform automated testing and calibration of the system.

Figure 2 shows a fragment of the detailed automated integration workflow for freight estimation that we have developed. The workflow for Figure 1 accesses 11 different sources and requires about 40

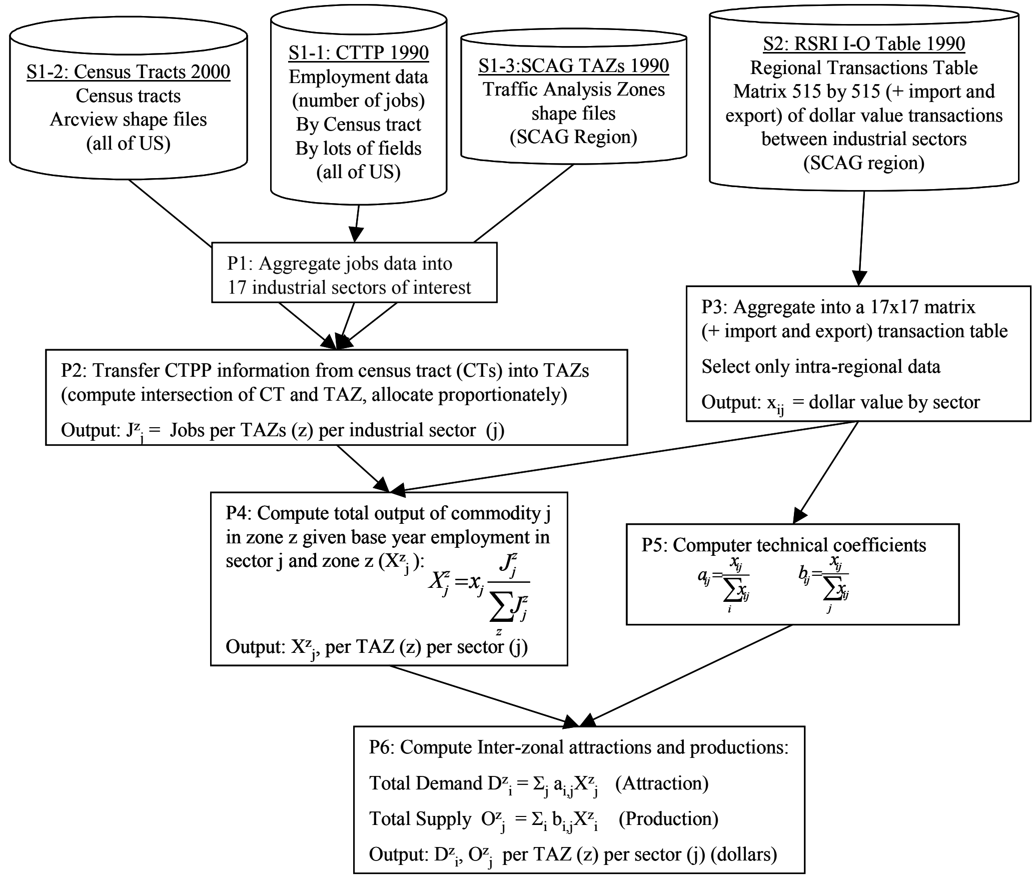

S1-2: Census Tracts 2000
Census tracts
Arcview shape files
(all of US)
S1-1: CTTP 1990
Employment data
(number of jobs)
By Census tract
By lots of fields
(all of US)
S1-3:SCAG TAZs 1990
Traffic Analysis Zones
shape files
(SCAG Region)
S2: RSRI I-O Table 1990
Regional Transactions Table
Matrix 515 by 515 (+ import and
export) of dollar value transactions
between industrial sectors
(SCAG region)
P1: Aggregate jobs data into
17 industrial sectors of interest
P2: Transfer CTPP information from census tract (CTs) into TAZs
(compute intersection of CT and TAZ, allocate proportionately)
Output: J^z_j = Jobs per TAZs (z) per industrial sector (j)
P3: Aggregate into a 17x17 matrix
(+ import and export) transaction table
Select only intra-regional data
Output: x_{ij} = dollar value by sector
P4: Compute total output of commodity j
in zone z given base year employment in
sector j and zone z (X^z_j):
$X^z_j = x_j \frac{J^z_j}{\sum_z J^z_j}$
Output: X^z_j, per TAZ (z) per sector (j)
P5: Computer technical coefficients
$a_{ij} = \frac{x_{ij}}{\sum_i x_{ij}}$
$b_{ij} = \frac{x_{ij}}{\sum_j x_{ij}}$
P6: Compute Inter-zonal attractions and productions:
Total Demand $D^z_i = \Sigma_j a_{i,j} X^z_j$ (Attraction)
Total Supply $O^z_j = \Sigma_i b_{i,j} X^z_i$ (Production)
Output: D^z_i, O^z_j per TAZ (z) per sector (j) (dollars)

Figure 2. Fragment of integration workflow: access to geo-spatial (S1-2, S-3), web (S1-1), and database (S2) sources, and processing steps (Pi).

processing steps. Figure 2 focuses on the computation of inter-zonal attractions and productions (corresponding to the dashed area in Figure 1). The data necessary for this calculation are obtained from four sources. The first source is the Census Transportation Planning Package (S1-1 in Figure 2) that provides employment data by census tract (CT) by place of work for the major industrial sectors. In order to integrate this information with other available sources, we need to aggregate the employment activities by 17 industrial sectors of interest (operation P1). Similarly, the spatial unit of our analysis is the Traffic Analysis Zone (TAZ). Therefore, we need to compute employment figures for each TAZ. For this, we need to access to two additional sources, S1-2 and S1-3, which provide the spatial descriptions of the Census tracts and SCAG TAZs, respectively. Then, we perform a change of zoning system (P2), a spatial operation that involves computing the intersection of CTs and TAZs and allocating the employment data proportionally to the area of the intersections.

Simultaneously, we access the input-output model corresponding to the SCAG region from the Regional Science Research Institute (source S2). This is a transactions table for 515 industrial sectors. Again, we aggregate the detailed industrial sectors into a coarser set of 17 sectors and select only intra-regional data (operation P3), obtaining a 17 by 17 matrix.

Combining the results of operations P2 and P3, we can compute the output of each sector within a TAZ (operation P4). We allocate the commodity output of each industrial sector within each TAZ proportionately to the jobs in the same sector. From these, we can compute the total attractions and productions of each TAZs by the 17 industrial sectors of interest (operations P5 and P6).

The workflows in Figures 1 and 2 illustrate the desirability of accessing, integrating, and processing data. We are currently developing a framework to more easily specify workflows and automate this process. In particular, we are exploring using an extension to Heracles (Knoblock et al 2001), a constraint-based integration system. Heracles can handle geo-spatial and traditional databases and web sources. Heracles integrates the data in these sources and manipulates the data using a network of constraints. The constraints can implement, in addition to source calls, arbitrary computational components. Thus, we plan to use Heracles constraints to implement the different processing steps.

4 Discussion

We have presented current work on developing an automated integration system for freight flow analysis and planning. This project will allow analysis and comparison of different scenarios, including the impacts of expanded international trade, the impacts of increased highway or facilities congestion, the contribution of trucking to highway congestion, the relationship between employment location and commodity flows, etc. Although the project focuses initially in the Los Angeles metropolitan area, we expect that transferring the results to other regions will be straightforward. The computational framework will be the same, many data sources have a national scope, so they can be reused. We would only need to add sources specific to a particular area. In summary, we hope that our system will provide a tool that will allow regional planners and policymakers to make better and more informed decisions.

To ensure that the results of this project are relevant to their intended users, we have assembled an advisory group of transportation planning experts from local, regional, and state agencies, including the Los Angeles County Metropolitan Transportation Authority, the Ports of Long Beach and Los Angeles, the Southern California Association of Governments, and California Department of Transportation.

References

Giuliano, Genevieve (1998), Information Technology, Work Patterns and Intra-metropolitan Location: A Case Study. Urban Studies, Volume: 35 Number: 7, 1077–1095

Gordon, Peter and Qisheng Pan (2001) Assembling and Processing Freight Shipment Data: Developing a GIS-Based Origin-Destination Matrix for Southern California Freight Flows. National Center for Metropolitan Transportation Research (www.metrans.org)

Gordon, Peter; Liao, Yu-chun and Richardson, Harry Ward, (1998*) Household commuting: implications of the behavior of two-worker households for land-use/transportation models.* Network infrastructure and the urban environment : advances in spatial systems modeling. Berlin; New York : Springer.

Graham, Stephen and Simon Marvin, (1996) *Telecommunications and the city: electronic spaces, urban places*. London ; New York : Routledge.

Hedges, CA (1971), *Demand forecasting and Development of a Framework for Evaluation of Urban Commodity flow: Statement of the Problem.* Special Report 120: Urban Commodity Flow, pp 145-148. Highway Research Board, Washington DC.

Craig A. Knoblock, Steve Minton, Jose Luis Ambite, Maria Muslea, Jean Oh, and Martin Frank (2001). *Mixed-Initiative, Multi-source Information Assistants*. The Tenth International World Wide Web Conference (WWW10), Hong Kong.

List, GF and MA Turnquist (1994). *Estimating truck travel patterns in Urban areas,* Transportation Research Record 1430, 1-9.

List, GF., V Papayanoulis, LA Konieczny and CL Durnford. (2002), *A Best practice truck flow estimation model for the New York City Region,* TRB 2002 Annual meeting.

Sheffi, Y. (1985) *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. New Jersey: Prentice Hall.

Southern California Association of Governments (SCAG). (1999) *SCAG Heavy Duty Truck Model and SCAG VMT Estimates*. Los Angeles, California.

US Census Bureau. (1997) *Commodity Flow Survey*. Washington DC: U.S. Department of Transportation and U.S. Department of Commerce.

US Department of Transportation (USDOT) (2000) *National Transportation Statistics 2000,* Table 1-41, p. 64.

US Department of Transportation (USDOT) (1996) *Quick Response Freight Manual –Final Report*. Washington, D.C.

US Department of Transportation (DOT), (1998), *US Freight: Economy in Motion*, p.6.

Wilson, AG (1970). Entropy in Urban and Regional Modeling, Pion, London.

Willumsen, LG (1978). *OD Matrices from network data: a comparison of alternative methods for their estimation*, Proceedings of the PTRC Summer annual meeting: 1978 Seminar in transportation models, PTRC educational research services ltd. London.

Willumsen LG (1984). *Estimating time dependent trip matrices from traffic counts,* Ninth International symposium on transportation and traffic theory, VNU Science press, Utretch, The Netherlands