INTEGRATING HETEROGENEOUS DATA SOURCES FOR BETTER FREIGHT FLOW ANALYSIS AND PLANNING

Submission Date: 7/29/02

Revised 11/14/02

Author 1
Jose-Luis Ambite
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina Del Rey, CA 90292
Ph 310-448-8472
F 310-822-0751
Email ambite@isi.edu

Author 2 (Contact Author)
Genevieve Giuliano
School of Policy, Planning and Development
University of Southern California
Los Angeles, CA, 90089-0626
P 213-740-3956
F 213-740-0001
Email giuliano@usc.edu

Author 3
Peter Gordon
School of Policy, Planning and Development
University of Southern California
Los Angeles, CA 90089-0626
P 213-740-1467
F 213-740-6170
Email Gordon@usc.edu

Author 4
Qisheng Pan
School of Policy, Planning and Development
University of Southern California
Los Angeles, CA 90089-0626
P 213-740-1467
F 213-740-6170
Email qpan@usc.edu

Author 5
Sandipan Bhattacharjee
School of Policy, Planning and Development
University of Southern California
Los Angeles, CA, 90089-0626
P 213-821-0780
F 213-740-0001
Email sandipab@usc.edu

ABSTRACT

We present ongoing work on developing an automated data integration system for intra-metropolitan freight flow analysis and planning. To overcome the limitations of current estimation methods for commodity flows, we use reliable secondary sources, including small-area employment data, and derive estimates in a plausible 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 from many sources and enhance the ability to explore different scenarios.

INTEGRATING HETEROGENEOUS DATA SOURCES FOR BETTER FREIGHT FLOW ANALYSIS AND PLANNING

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 1999 for example, of the total bill of $562 billion the US spent on freight transportation, trucks carried 81%, i.e. $457 billion (1). 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 (2). By some measures, U.S. commodity trade with the rest of the world is growing even faster.

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 and often conflicts with demands for passenger commuter service. Increased truck traffic has accelerated deterioration of highways. In addition, rapid changes in economic linkages are leading to ever changing flow patterns and the spatial restructuring of metropolitan areas (3,4,5).

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 summary of recently completed research 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 intra-metropolitan 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. The second section describes current freight flow methods and their problems. The third section presents the conceptual framework for our model. We discuss the motivations for our "bottom-up" approach and describe the model as developed to date. We present our plans for constructing and testing a continuously updatable freight flow estimation model in the fourth section. The last section presents some results from applying the model, and the final section discusses conclusions.

CURRENT METHODS FOR FREIGHT FLOW ESTIMATION

Urban researchers have a limited understanding of intra-metropolitan freight flows, both because passenger transport has been the dominant concern and because of the dearth of detailed freight flow data. 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.

There is an extensive literature on urban transportation network modeling, and the state-of-practice is well advanced (6,7,8,9). Transportation network models estimate vehicle flows on the network as a function of the spatial distribution of economic activity, demand for spatial interaction, transport prices and network capacity. A variety of theoretical models have been developed, most of which focus exclusively on urban passenger movement. Goods movement is typically modeled as an inter-regional trade problem, with trade a function of local production costs, transport costs, and commodity demand (e.g. 10, 11).

The state of practice lags behind state of the art theoretical models, because of the complexity of the system being modeled and the level of detail required for planning purposes. The most commonly used "4-step model" does not consider goods movement. Strategies for incorporating goods movement include 1) using fixed factors based on passenger vehicle flows and observed truck counts at a small number of locations on the highway network; 2) developing a truck origin-destination matrix from survey data and assigning truck trips to the network; or 3) using goods movement trip generation models and following a 4-step procedure to produce truck flows on the network (e.g. 12, 13). Rail freight flows are not usually modeled at the metropolitan level.[1] The first strategy is very approximate; the second requires detailed origin-destination studies that are expensive and often resisted by shippers. The third strategy requires commodity specific trip data, also difficult and costly to obtain and update at the intra-metropolitan level.

The growing importance of freight in urban transportation calls for better freight modeling methodologies. An improved model should: (a) have solid behavioral foundations, (b) be multi-modal, (c) be able to analyze interactions between passenger and freight trips and (d) be able to take feedback from policy changes (14).

Data Problems

Current methods of intra-metropolitan freight flow estimation are limited by lack of data that is appropriate, accessible, and reliable. Ideally, one would like to have current and 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 collecting 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. Metropolitan level analysis requires fine geography; most existing data are at a 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 and time intervals. While we cannot solve all the problems, we have developed ways to circumvent many of them.

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 (12). Truck trip generation rates were estimated from the number of jobs in the employment sectors 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 easy 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 were assigned to regional highway networks, and VMTs were estimated for emissions analyses (13). 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 its Intermodal Transportation Management System (ITMS) for statewide transportation planning since 1996. ITMS estimated freight movement by different modes based on data from the Surface Transportation Board's Carload Waybill Sample, the Reebie Associate's Transearch database, the Dri/McGraw Hill's US economy forecast, and the 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.

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. We use a bottom-up approach that makes use of 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 (15; 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 US in terms of container shipments, as well as the Los Angeles International airport among others. The problem is to generate a matrix of intra-regional commodity flows that includes both intra-regional economic activity and import/export activity.

Small area employment data and a regional input-output model form the basis for generating intra-regional commodity flows. The approach begins with an input-output model of the local economy and divides transactions into two commodity flow types, intra- and inter-regional; a transactions table less interregional and international 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 freight trips producxed and freight trips attracted by major commodity for each of the region's 1527 Traffic Analysis Zones (TAZs) intra-regional supply and demand. Summing, they get intra-regional shipments produced and shipments attracted for each zone by sector.

Specifically,. Equation (1) tallies the total commodity i required to support production in zone z:

$$D^z_i = \Sigma_j a_{i,j} \bullet X^z_j \quad (1)$$

where X^z_j = the total output of commodity j in zone z given base year employment in sector j and zone z, and

$a_{i,j}$ = the i, jth element of **A**, the matrix of value demand coefficients for the (open) input-output model. This is the flow from i to j per unit output of j.

Total shipment attractions (destinations) is the sum of commodities demanded from all other zones, for all industries in the zone to accommodate local final demand not associated with households. This includes commodities demanded from transshipment zones (imports).

Similarly, the total supply of output j furnished by zone z is calculated in equation (2),

$$O^z_j = \Sigma_i b_{i,j} \bullet X^z_i \quad (2)$$

where X^z_i = the total output of commodity i in zone z given base year employment in sector i and zone z, and

$b_{i,j}$ = is the i, jth element of **B**, the matrix of value supply coefficients for the (open) input-output model. This is the flow from i to j per unit output of i.

Total shipment productions (origins) are the sum of outputs of all commodities produced by all industries in the zone and shipped to all other zones to accommodate local final demand

not associated with households. Again, this includes commodities demanded by transshipment zones.

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 US Commodity Flow Survey (16), 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 and international trade. This calculation only concerns a limited number of zones: for the Los Angeles area, the two major seaports (Long Beach and Los Angeles); the five major airports involved in freight shipping; three major rail yards and six highway entry-exit points. These have jobs associated with both intra- and inter-regional trade. The former are estimated in Equations (1) and (2).

The estimated inter-regional trade data are also useful for updating and checking the ITMS data, which theoretically include all 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 and international trade. Los Angeles International Airport (LAX) and (to a lesser extent) Ontario International Airport reported detailed data on international shipments. National shipment data were not as readily available. In the Gordon-Pan application, some rule-of-thumb proportions from a SCAG study were utilized. 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. SCAG's 1996 Interregional Goods Movement Study provided total tonnage of shipments in and out of the region by mode, as well as the ratios of goods originating in or destined for the SCAG region, based on a one-time data collection. The completed origin-destination matrix implied 6.21 million metropolitan area jobs, a sum reasonably close to the actual 1990 figure of 6.61 million jobs, considering the limited data sources used in the first approximation. Further work will replace many of these with more plausible sources.

Once the origin-destination matrix of freight flows was estimated, a traffic assignment model was used to assign these flows to region's highway network. Freight flows (in passenger-car equivalents) were *added* to SCAG's estimated passenger flows on all links. The network assignment was performed using the Sheffi User-Optimal-Strict model (17)

In summary, Gordon and Pan developed a methodology for estimating the detailed commodity flows based on secondary sources. They used plausibly reliable secondary sources, as known to them at the time , and derived the estimates in a more reasonable way than the competing approaches discussed previously. 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.

AUTOMATED INTEGRATION AND ESTIMATION OF FREIGHT FLOWS

Our current work, to be completed in the next stage of the rsearch, 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 one 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.

The first task of our work was to develop a detailed workflow of the Gordon and Pan model. Figure 2 shows a fragment of the workflow we have developed. The workflow for Figure 1 accesses 11 different sources and requires about 40 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 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 compute employment figures for each TAZ. For this, we access 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 IMPLAN[2]. 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 TAZ 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 these sorts of workflows to automate the process. In particular, we are exploring extensions to two of the integration systems that have been developed at the Information Agents group at USC/ISI (http://www.isi.edu/info-agents/).

The first system is Theseus (18, 19), an efficient plan execution system for information management agents. Theseus provides an expressive language to define data gathering and processing plans. It provides operators for data access from local and remote data sources, relational (select, project, join, etc.) and XML data manipulation (using XML Query (20)), data manipulation, aggregation, and user-defined functional application. Plans can be named and

[2] IMPLAN is available for use in any U.S. metro area

resued as subplans. Recursive invocation of subplans is allowed. Theseus supports a number of capabilities that enable practical information management, including repeated and periodic query execution, conditional plan declarations, query result aggregation, and flexible communication of results. The Theseus plan executor focuses on efficiency, with support for data pipelining, and dataflow-based event-driven parallel execution.

We used Theseus to perform a proof-of-concept implementation of the part of the workflow of Figure 1 that gathers and derives freight data for the five major airports in the Los Angeles region. The air cargo data are available as an HTML page at the Los Angeles World Airports web site (http://www.lawa.org). We created a wrapper for the site that extracts the most recent web data live at run-time. A wrapper is a program that provides access to a semi-structures web source, such as HTML site, as if it were a database. This task was greatly facilitated by using our tools for semi-automatically generating wrappers (21). Our tool automatically generates a wrapper by learning data extraction rules for a site from a few user-provided samples of the desired data. Since the web site only provides data for Los Angeles International Airport (LAX), we used information about the percentage of total freight that each airport uses (also found at the LAWA web site) to estimate the air freight at the remaining airports. Similarly, we used an estimate of the dollar cost per ton of freight to calculate the dollar value of the freight at each airport. These percentages and other parameters were stored in a local database. The computational steps and assumptions were implemented as Theseus operators and the workflow as a Theseus plan. We deployed the Theseus plan for airport cargo estimation as a Java servlet, so that calling the plan and displaying results can be done from any web browser. The demo can be accessed at http://chauvet.isi.edu:8888/servlets/index.html.

The second system is Heracles (22), a constraint-based integration system. Heracles integrates the data from multiple sources and manipulates the data using a network of constraints. Heracles handles geospatial and traditional databases and web sources. The constraints have the capability to implement arbitrary computational components as well as calls to data sources. In addition, Heracles provides a user interface, automatically generated from the specification of the constraints, that allows the user to interact with the system by entering values for some of the variables or choosing among the computed values. This would allow for interactive what-if analysis. We are currently developing a new version of Heracles with additional geospatial integration capabilities that would be suited to handle geospatial data access and manipulation such as required in the workflows of Figures 1 and 2.

EMPIRICAL RESULTS

The modeling process described above makes it possible to estimate freight commodity flows by mode on the transportation network. Rail and air network flows are straightforward; the more challenging problem is to estimate truck flows on the highway network. We have run the model and estimated truck flows for the SCAG region, and now briefly discuss our results.

We start with the commodity flow data in dollar value and tonnage, which is separated into intraregional attractions by the 1527 TAZs, and interregional inbound and outbound freight flow at each major export node (airport, seaport, railyard, and highway entry point) in the region as described previously. The intra-regional and inter-regional commodities are combined in terms of the geographic locations of the TAZ and the major export nodes; then the freight tonnage is converted into average daily heavy duty truck (HDT) equivalent units.

The HDTs are distributed using a gravity model. The distance decay coefficient values for freight trips are calibrated to minimize the difference between the “observed” and

"estimated" freight trip productions (23, 15). Coefficients are calculated for four freight sectors. These are applied to the HTD trips associated with commodity flows for the given sector. The result is a 1533 by 1533 (1527 SCAG TAZs plus 6 external zones as highway entry/exit points) HDT O-D trip table.

The assignment of HDTs to the highway network takes as given the equilibrium assignment of passenger car trips (e.g. there is already congestion, and the passenger car assignment does not change). The traffic assignment submodel performs a 3-hour AM peak assignments for passenger car equivalents (PCEs), hence the HDT O-D trip table must be converted to PCEs and factored down to the peak hour. The HDT traffic assignment was performed using the Sheffi (17) UO-S approach described earlier.

In order to assess the plausibility of the traffic assignment results, it is desirable to compare estimated HDT volumes with actual volumes. We were able to use actual count data collected by SCAG as part of its 1994 Heavy Duty Truck Model study.[3] Eleven regional screenlines were established for the study, and 24 traffic counts were conducted for all freeways, state highways and major arterials crossing each screenline. Figure 3 gives the location of the 11 screenlines.

Table 1 gives the 1994 actual count data by screenline. The first panel gives autos, HDTs, total vehicles, and the HDT share. The second panel gives PCEs for the trucks and HDT share of total PCEs. The heavy truck share ranges from 3.88 percent (screenline 4) to just over 8 percent (screenlines 6, 11). When HDTs are converted to PCEs, the HDT share increases proportionately.

We compare our HDT estimates with the SCAG's reports of actual counts in Table 2. We convert our model results to 24 hour PCEs using the same conversion factors as SCAG and compare them to SCAG 24 hour PCEs. Our model estimates are generally lower than actual, in some cases substantially lower (screenline 7). However, there is a significant correlation between the estimated and actual (0.72), and a simple linear regression of actual on estimated values gives R^2 of 0.54. See Figure 4.

In addition to all of the assumptions and data problems associated with our first cut at generating the commodity flow matrix, there are other problems associated with the traffic assignment itself that would account for differences: 1) our HDT estimate is based largely on 1996 data; for example, all the interregional commodity flow data collected for airport, seaport, rail, and highway are in 1996; 2) the passenger car assignment used 1990 data, the only available data source; 3) no attempt was made to calibrate the traffic assignment. Yet, overall we are encouraged by these results. Our very approximate estimate yields truck volume estimates that are reasonably close to observed values

DISCUSSION

We have presented a summary of current work on developing an automated integration system for freight flow analysis and planning in order to highlight the promise of continued work to complete automation, tested with various data sources and the analysis and comparison of 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 focused on the Los Angeles metropolitan area, we expect that transferring the results to other

[3] SCAG defines HDT as vehicles with gross vehicle weight greater than 8500 pounds.

regions will be straightforward. It could be argued that Los Angeles is an extreme test for the model, given its size, diversity, and volume of international trade. The computational framework will be the same, many data sources have a national scope, so they can be utilized. We would only need to add sources specific to a particular area. In summary, we expect 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. Although the model does not yet meet requirements for actual planning applications, the members have show substantial enthusiasm for our work and are available to offer continued advice.

Acknowledgement
This research was supported by the National Science Foundation and by the METRANS Transportation Center through grants supported by the USDOT and the California State Department of Transportation. Comments on a previous version of the paper by members of the project's advisory group, by David Forkenbrock, Morton O'Kelly, and by TRB anonymous referees are appreciated. All errors and omissions are the responsibility of the authors.

REFERENCES

1. US Census Bureau (2001) *Statistical Abstract of the United States, 121st edition.* Washington, DC. Table 1051.
2. US Department of Transportation (USDOT) (2002) *National Transportation Statistics 2001,* Table 1-41, Washington, DC.
3. Graham, Stephen and Simon Marvin, (1996) *Telecommunications and the city: electronic spaces, urban places*. London ; New York : Routledge.
4. 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.
5. Giuliano, Genevieve (1998), Information Technology, Work Patterns and Intra-metropolitan Location: A Case Study. Urban Studies, Volume: 35 Number: 7, 1077–1095
6. Wilson, AG (1970). Entropy in Urban and Regional Modeling, Pion, London.
7. List, GF and MA Turnquist (1994). *Estimating truck travel patterns in Urban areas,* Transportation Research Record 1430, 1-9.
8. 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.
9. 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.
10. Harker, P (1987) *Predicting Intercity Freight Flows*, VNU Science Press, Utretch, The Netherlands.
11. Ogden, KW (1992) *Urban Goods Movement: A Guide to Policy and Planning.* Ashgate Publishers, Aldershot, Hants.
12. US Department of Transportation (USDOT) (1996) *Quick Response Freight Manual –Final Report*. Washington, D.C.
13. Southern California Association of Governments (SCAG). (1999) *SCAG Heavy Duty Truck Model and SCAG VMT Estimates*. Los Angeles, California
14. 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.
15. 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)
16. US Census Bureau. (1997) *Commodity Flow Survey*. Washington DC: US Department of Transportation and US Department of Commerce.
17. Sheffi, Y. (1985) Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods. New Jersey: Prentice Hall.
18. Greg Barish, Dan DiPasquo, Craig A. Knoblock, Steven Minton (1999) "An Efficient Plan Execution System for Information Management Agents" ACM CIKM Web Information Data Management Workshop. Kansas City, MO, USA.
19. Greg Barish and Craig A. Knoblock (2002) "An Expressive and Efficient Language for Information Gathering on the Web", Proceedings of the Sixth International Conference on AI

Planning and Scheduling (AIPS-2002) Workshop: Is There Life Beyond Operator Sequencing? - Exploring Real-World Planning, Toulouse, France. April 2002.
20. Boag, S, D. Chamberlin, M. Fernandez, D. Florescu, J. Robie, J. Simeon (2002) "Xquery 1.0: An XML Query Language," WC3 Working Draft, http://www.w3.org/TR/xquery/.
21. Craig Knoblock, Kristina Lerman, Steve Minton, Ion Muslea (2000) "Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach", IEEE Data Engineering Bulletin, 23(4):33-41
22. 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.
23. Cho, Sungbin et. al. (1999). *Integrating Transportation Network and Regional Economic Models to Estimate the Costs of a Large Earthquake: NSF Draft Report*. Los Angeles, California.

TABLES AND FIGURES

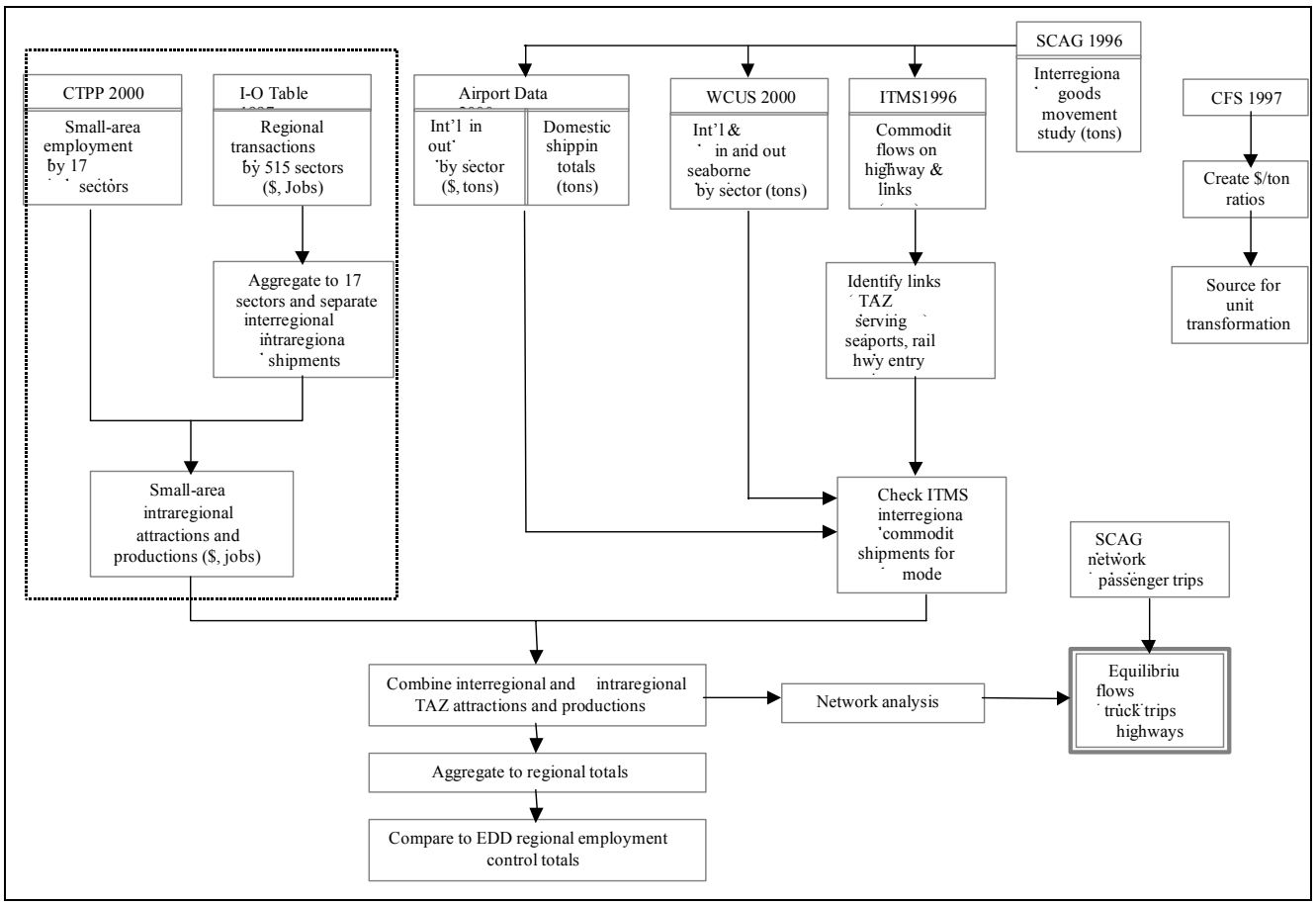

Figure 1. Summary of Freight Data Collection and Processing.
(The dashed area is shown in detail in Figure 2)

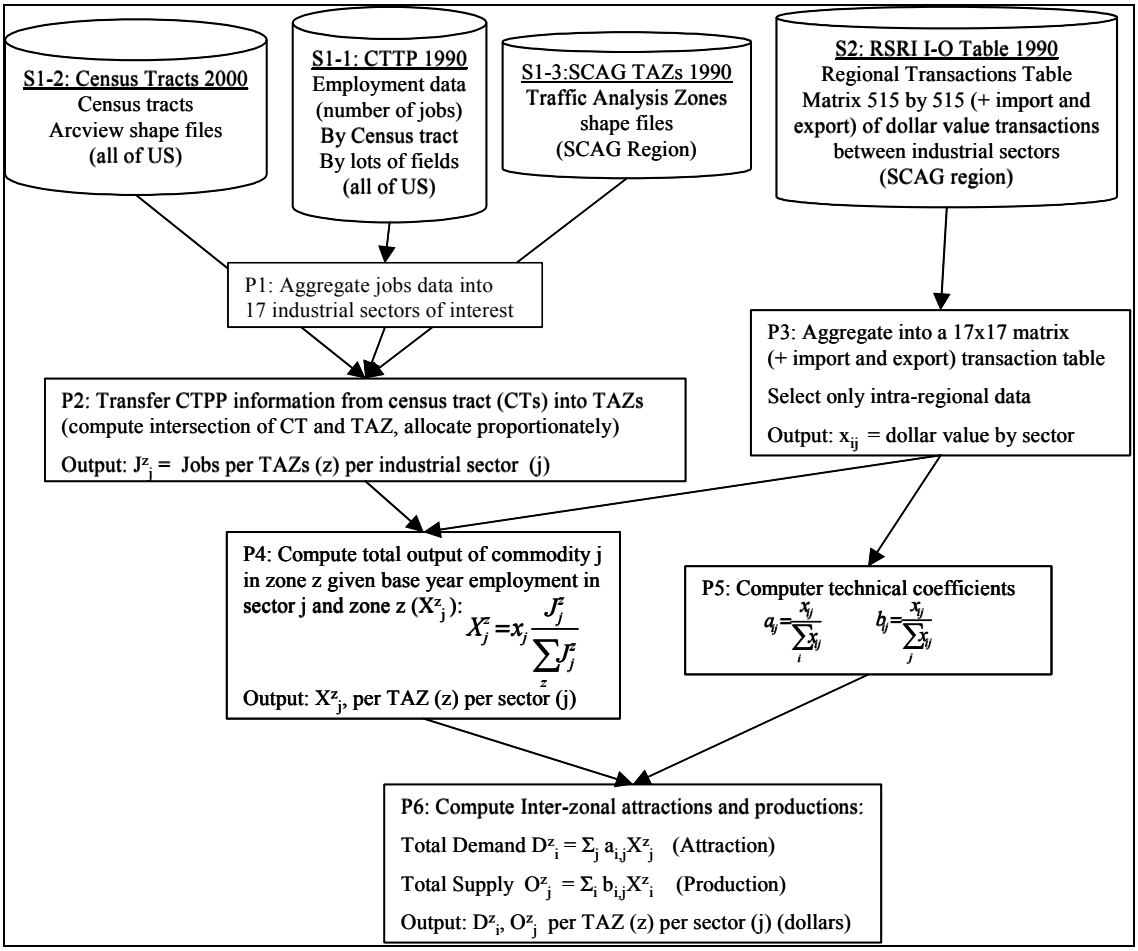

Figure 2. Fragment of integration workflow: access to geo-spatial (S1-2, S-3), web (S1-1), and database (S2) sources, and processing steps (Pi).

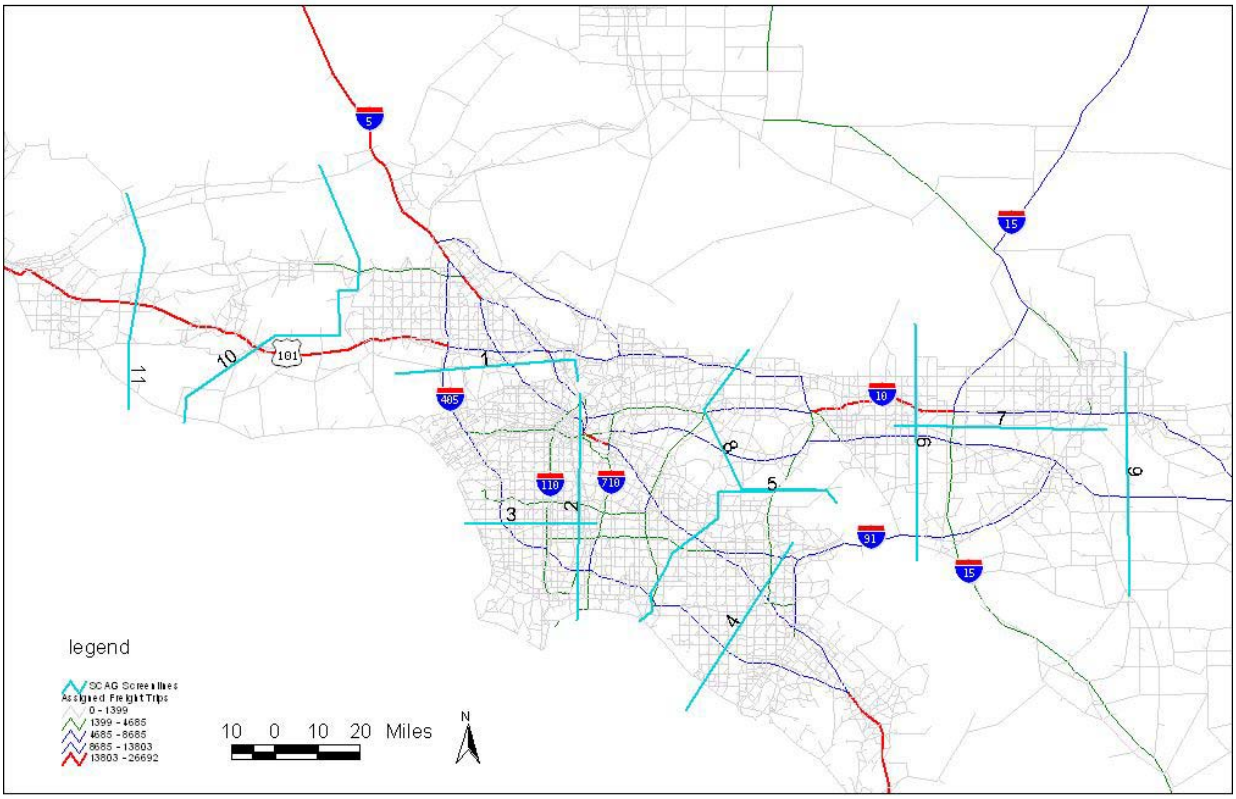

Figure 3 SCAG 1994 Modeling Screenlines

Table 1. SCAG 1994 Traffic Counts on Regional Screenlines[a]

SCAG	24-hr SCAG 1994 Ground Counts (ADT)				PCE Value[b]	PCE(ADT)[c]	
Screenlines	Autos	HDT	Total	HDT/Total		HDT	HDT/Total
1	1,442,833	61,870	1,504,703	4.11%	3.9	243,355	16.17%
2	2,210,602	106,041	2,316,643	4.58%	3.9	417,095	18.00%
3	1,179,505	59,381	1,238,886	4.79%	3.2	191,999	15.50%
4	1,618,517	65,344	1,683,861	3.88%	3.9	257,020	15.26%
5	1,425,686	84,261	1,509,947	5.58%	3.2	271,040	17.95%
6	836,320	73,546	909,866	8.08%	2.9	210,832	23.17%
7	617,170	52,893	670,063	7.89%	3.6	188,652	28.15%
8	1,183,291	84,400	1,267,691	6.66%	3.2	271,487	21.42%
9	363,082	29,135	392,217	7.43%	3.6	103,915	26.49%
10	379,735	20,495	400,230	5.12%	6.7	136,633	34.14%
11	179,632	15,762	195,394	8.07%	8.9	139,756	71.53%

[a]Source: SCAG Heavy Duty Truck Model and VMT Estimation Final Report (SCAG 1999)
[b]Note: PCE values were estimated using Table 18 of the SCAG Heavy Duty Truck Model and VMT Estimation Final Report (SCAG 1999)
[c]Note: HDT ADT (Average Daily Traffic) was converted into PCE/day using the estimated PCE value

Table 2: Comparison of Model Estimates and 1994 Actual HDT, 24 hr, PCEs

Screenline	'94 Actual	Model	Difference	% Difference
1	243,355	222,639	-20,716	-9%
2	417,095	299,786	-117,309	-28%
3	191,999	71,524	-120,475	-63%
4	257,020	169,547	-87,473	-34%
5	271,040	203,560	-67,480	-25%
6	210,832	217,482	6,650	3%
7	188,652	28,737	-159,915	-85%
8	271,487	197,476	-74,010	-27%
9	103,915	129,089	25,174	24%
10	136,633	132,738	-3,896	-3%
11	139,756	123,152	-16,604	-12%

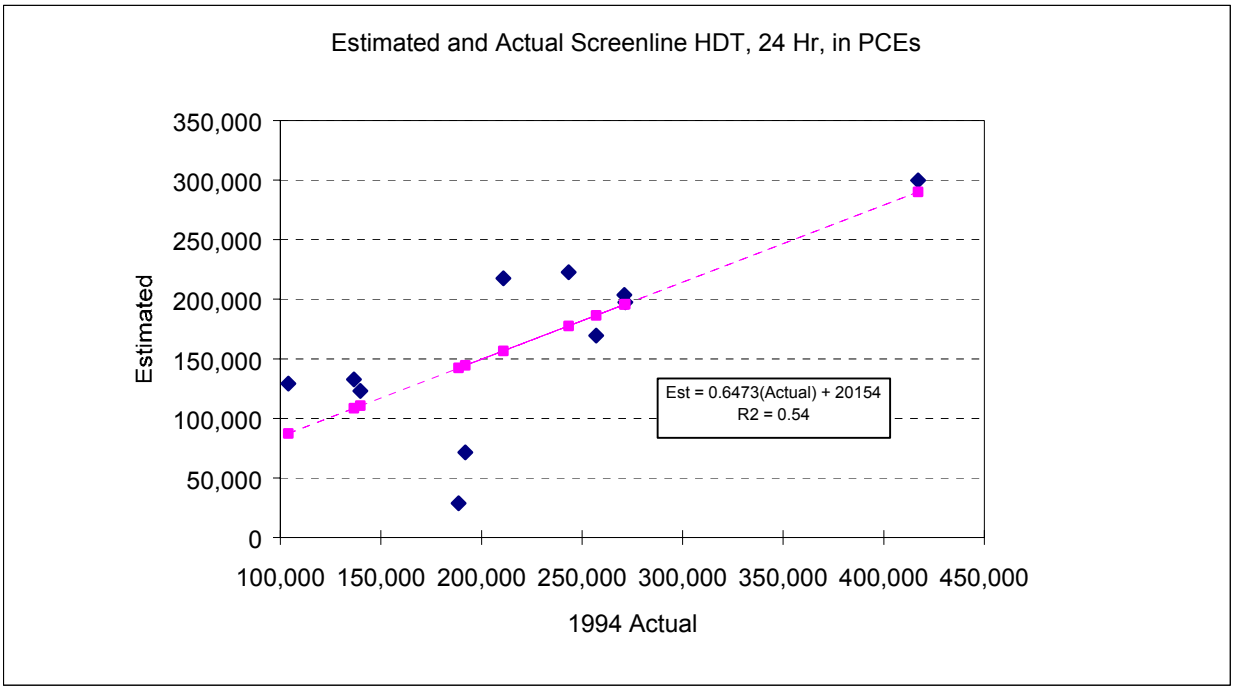

Figure 4 Model and Actual Screenline Results