Geographic Scalability and Supply Chain Elasticity of a Structural Commodity Generation Model Using Public Data

UCI-ITS-WP-12-4

Fatemeh Ranaiefar
Joseph Y.J. Chow, Ph.D.
Daniel Rodriguez-Roman
Pedro V. Camargo
Stephen G. Ritchie, Ph.D

Institute of Transportation Studies
University of California, Irvine; Irvine, CA 92697-3600, U.S.A.
fatemeh.ranaiefar@gmail.com, joseph.chow@gmail.com,
drr.caam@gmail.com, pcamargo@uci.edu, sritchie@uci.edu

October, 2012

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

Geographic scalability and supply chain elasticity of a structural commodity generation model using public data

Fatemeh Ranaiefar
Ph.D. Candidate
Institute of Transportation Studies
University of California
Irvine, CA, USA, 92697
franaief@uci.edu

Joseph Y.J. Chow
Assistant Professor
Department of Civil Engineering
Ryerson University
Toronto, ON, Canada, M5B 2K3
joseph.chow@ryerson.ca

Daniel Rodriguez-Roman
Ph.D. Student
Department of Civil and Environmental Engineering
Institute of Transportation Studies
University of California
Irvine, CA, USA, 92697
d.rodriguez@uci.edu

Pedro V. Camargo
Ph.D. Student
Institute of Transportation Studies
University of California, Irvine, CA, USA, 92697
pcamargo@uci.edu

Stephen G. Ritchie
Professor
Department of Civil and Environmental Engineering
Institute of Transportation Studies
University of California
Irvine, CA, USA, 92697
sritchie@uci.edu

Revision Submitted: October 31, 2012
Word Count: 5,694
Tables and Figures: 4 Tables + 4 Figures = 2000 words
Total Word Count: 7,694

Accepted in Proceedings of the TRB 92nd Annual Meeting
Submitted for Publication review for the Transportation Research Record

ABSTRACT
Freight forecasting models are data intensive and require many explanatory variables to be accurate. One problem, particularly in the United States, is that public data sources are mostly at highly aggregate geographic levels, while models with more disaggregate geographic levels are required for regional freight transportation planning. Second, supply chain effects are often ignored or modeled with economic input-output models which lack explanatory power. This study addresses these challenges by considering a structural equation modeling approach, which is not confined to a specific spatial structure as spatial regression models would be, and allows for correlations between commodities. A FAF-based structural commodity generation model is specified and estimated and shown to provide a better fit to the data than independent regression models for each commodity. Three features of the model are discussed: indirect effects, supply chain elasticity, and intrazonal supply-demand interactions. A validation of the geographic scalability of the model is conducted using data imputed with a goal programming method.

Keywords: freight, forecasting, structural equations, commodity generation, supply chain, regression

1 INTRODUCTION

Regional commodity-based freight forecasting research has gained steady traction in the last fifteen years. In the United States, national *[1]* and statewide freight modeling studies *[2]* are now required in transportation plans. A number of innovations in advanced freight forecasting have been developed (*[3]*, *[4]*) to address these needs. However, data sources for these models are difficult to obtain due to their high costs and the proprietary nature of the private sector *[5]*. Existing public data sources tend to be at highly aggregate geographic levels, beyond what would be useful for regional freight transportation planning.

Freight modeling researchers have dealt with this issue in two ways. The first is to forego the use of public data in favor of expensive, firm-level (or highly disaggregate zones) data for freight models. Examples range from integrated land use models (*[6]-[8]*) to agent-based simulation models (*[9]-[12]*).

The second approach is to make the most use of publicly available data to produce aggregate freight models. Many of the statewide models in the U.S. fall into this category *[2]*. Freight models in this direction tend to focus on ways of either maximizing the use of limited data or introducing sophisticated models. Giuliano *et al. [13]* and Anderson *et al. [14]* imputed data from multiple secondary sources. Novak *et al. [15]*, on the other hand, proposed a more sophisticated spatial regression model with a conventional public data source. In this study, we argue in favor of public data using both sophisticated modeling and data imputation techniques.

In order to address this challenge, one key point needs to be clarified. Many freight models rely on disaggregation of aggregate data *[16]* so that models can be estimated and applied at a finer level of geographic detail. However, the geographical disaggregation error ends up being bundled with other unobservable noise in the data. As stressed by Holguín-Veras *et al. [17]*, inconsistency between model structures from one geographical zone to another can be problematic. A more measurable approach is to estimate a model at a coarser geographic level and to define the model in such a way that it is scalable to different geographic aggregations. The structure remains consistent and the error from disaggregation is bundled with the model error instead, which can be quantified with fitness measures at both geographic levels.

This study presents an alternative modeling framework to capture relationships between commodities at coarse geographies with public data that can then be applied to finer geographies. We consider a structural equation modeling (SEM) framework that makes the most use out of available public data. A commodity generation model is developed along with the data preparation necessary to run the model.

The remainder of this study is organized as follows. Section 2 is a literature review that highlights freight modeling with public data in the last few years. Section 3 focuses on the data preparation and the application of a goal programming approach to impute data for county and sub-county zone forecasting. Section 4 describes the specification and fitness measures of the proposed SEM framework for commodity generation, and an analysis of the model elasticities. Section 5 presents a validation using California data as a case study. Section 6 is the conclusion.

2 LITERATURE REVIEW

The limitation of freight data is well documented (*[3]*, *[5]*). Unlike household travel survey data for passenger models, equivalent data at the same level of detail (i.e. firms) are prohibitively expensive to acquire. While vehicle-based data is more accessible in the form of GPS trajectories *[18]* or truck diaries *[19]*, commodity-based data is often limited to such highly aggregate geographies as the Commodity Flow Survey (CFS) in the U.S *[20]*.

The consequence is that regional freight forecasting models that rely on these public data sources cannot explain higher resolution geographical effects. Researchers have sought ways to overcome this challenge. Giuliano *et al. [13]* acknowledged this "data problem", which is not just an issue of data aggregation but the need to build a closed model under an environment of increasingly open and global trade and goods movement. For example, freight transportation in the United States spans multiple states and/or regions, so statewide or Metropolitan Statistical Area (MSA) freight models need to consider national freight flows as well *[21]*. Anderson *et al. [14]* also constrained themselves to using only public data sources to develop a model for Alabama. For more advanced model techniques to overcome this data limitation, Ben-Akiva and de Jong *[22]* considered a hybrid modeling framework that blended aggregate and disaggregate methods.

Commodity generation models are forecast tools that estimate the amount of commodities produced or consumed at a zone or by a firm, typically based on socioeconomic data. They require commodity data such as the CFS or the imputed Freight Analysis Framework (FAF) data. As such, commodity generation models offer an opportunity to study ways of maximizing the use of coarse public data. An example of mutually exclusive commodity groups based on aggregations of 2-digit SCTG codes *[20]* is shown in TABLE 1. NCHRP Synthesis 298 *[23]* provides a comprehensive synthesis of freight generation.

TABLE 1. Commodity groups (Base on FAF 2007 data base)

Commodity group	2-dig SCTG	Total production (K ton) in U.S	% share of total
G1- Agriculture products	1-4	2,288,940	12.1%
G2- wood and paper products	26-29	739,761	3.9%
G3 Crude petroleum	16	836,581	4.4%
G4-Fuel and oil products	17,18,19	3,045,422	16.1%
G5- Gravel, sand and non metallic minerals	10-13	3,266,321	17.3%
G6- Coal and metallic ores	14-15	1,565,204	8.3%
G7- Food, beverage, tobacco products	5-9	937,853	5.0%
G8- Manufactured products	24,30,39,40,42,43	1,003,725	5.3%
G9- Chemical, pharmaceutical products	20-23	862,184	4.6%
G10- Nonmetal mineral products	31	1,392,666	7.4%
G11- Metal manufactured products	32-34	813,600	4.3%
G12- Waste material	41	1,324,523	7.0%
G13- Electronics	35,38	85,548	0.5%
G14- Transportation equipment	36-37	198,996	1.1%
G15-Logs and lumber	25	517,410	2.7%
Total	-	18,878,735	100%

Linear regression is the most widely used approach for freight generation modeling (*[24], [25]*). This approach has several drawbacks, however. Production and consumption is typically estimated independently for different commodities or groups of commodities. This independence assumption ignores the high correlations between different commodities due to supply chains and land use patterns. Novak *et al. [15]* pointed out the correlations present between productions and consumptions of different commodity groups in the CFS data.

Alternative methodologies have been proposed to overcome these issues. Bastida and Holguín-Veras *[26]* proposed a cross classification approach for urban truck trip generation. Like other classification and decision tree approaches, this method may result in better fitting models but may lose explanatory power. Novak *et al. [15]* and Chun *et al. [27]* proposed spatial regression to correct for spatial autocorrelation—a linear correlation or dependence among variables based on spatial proximity. However, spatial regression assumes a fixed spatial structure in order to characterize the relationships. The fixed structure prevents the model from scaling to different geographical zone sizes. For example, the model from *[15]* is calibrated on CFS zones, which in California consist of five regions, and cannot be used to explain commodity production and consumption within the 58 counties in the state.

Structural equation modeling (SEM) is a flexible linear-in-parameters multivariate statistical modeling technique that has gained acceptance in the travel behavior research community *[28]*. SEM is a more generalized form of linear regression that allows endogenous variables to serve as causal variables for other endogenous variables, and can identify unobservable factors called latent variables (hence the structure). There are different methods of estimating the parameters of these models, such as full information maximum likelihood estimation or three stage least squares estimation. SEM allows for both confirmatory and exploratory modeling, such that hypothesized causal relationships and correlations can be tested.

Nonetheless, very few freight models have been estimated based on this method. Jonnavithula *[29]* proposed an SEM-based direct demand model to estimate total commodity flows by mode for the state of Florida using proprietary TRANSEARCH zip code data. SEM was used in that context to capture the inter-dependencies among different freight modes, while keeping commodities independent.

Instead of capturing inter-dependencies between mode shares, SEM can be used to capture inter-dependencies between different commodity groups and productions and consumptions, effectively inferring the unobserved supply chain and land use relationships at an aggregate industrial level, much like economic input-output models. It is more flexible than spatial regression because it does not require a fixed spatial structure and can be applied to a different geographic resolution. For example, a model can be estimated based on FAF data, and can then be applied to a finer FAZ level defined at county or sub-county level. SEM is used as a confirmatory approach in this study: the structure design is hypothesized and the sample data is evaluated to confirm whether it fits the hypothesized design. Even if a good fitting structure naturally exists, it needs to be identified prior to the estimation.

3 DATA PREPARATION

Suppose that a structural commodity generation model is estimated using national FAF3 data. In order to apply it to a finer geographical level, the model would require input data at a finer level of detail. Using California as an example, a five zone aggregation would be too coarse for statewide freight analysis. Figure 1 shows the state of California with 96 freight analysis zones (FAZs) that are defined at approximately county or sub-county level. The FAZ boundaries were chosen to respect county and air basin boundaries that are important to analyzing statewide policies. The zones are delineated to have a maximum of 500,000 employment, 1.5 million population, and maximized homogeneity in land use within each FAZ. A spatial regression model calibrated at the FAF level cannot be applied to these finer geographic zones.

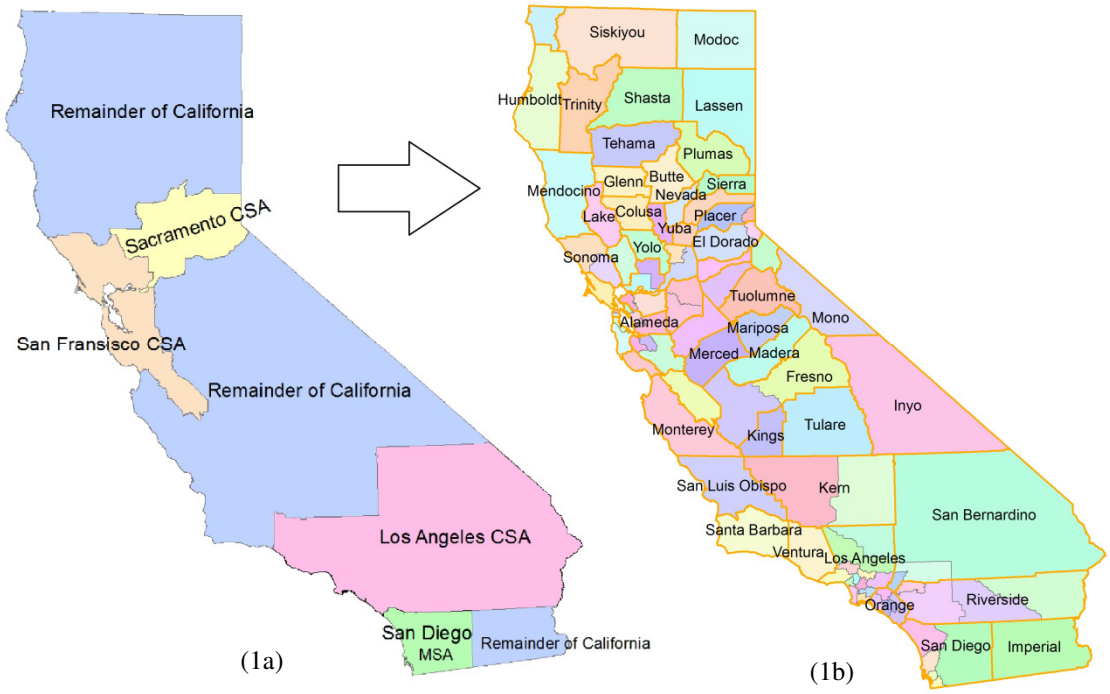


FIGURE 1. Five California FAF zones (1a) and 96 proprietary FAZs (1b).

Further disaggregation is not attempted because significant data suppression issues exist even at this level. Employment and number of establishment data at the county and state level by 3-digit North American Industry Classification System (NAICS) (hereafter 3-digit level) was obtained from the 2007 County Business Patterns (CBP*) [30]*. CBP provides employment by industry for up to 6 hierarchical levels. For confidentiality reasons, the number of employees in certain CBP entries is suppressed and substituted by a flag. A flag provides information on where the suppressed value lies (e.g., a flag means that the number of employees lies between 0 and 19). More detailed levels (4, 5 or 6 digits) have higher percentages of flags, whereas 3-digit industrial employment categories are compatible with 2 digit SCTG commodity grouping used in many freight models *([2],[16])*. Data suppression problems are also present in the Census of Agriculture (CoA) *[31]*, which was used to gather agriculture related variables, and in the Bureau of Economic Analysis' (BEA) Regional Economic Accounts, used to obtain manufacturing sector GDP *[32]*.

3.1 Data Imputation Methodology

Zhang and Guldmann *[33]* showed that midpoint approximation of suppressed data can lead to severe data inconsistencies, and proposed a goal programming approach to impute suppressed 2000 county-level CBP data. Their method was adapted for our study to impute the suppressed employment data for California CBP data at the 3-digit level. This methodology was selected since it ensures internal data consistency across geographical areas and industrial sectors. The following linear optimization problem in Equations (1) – (7) was solved.

$$\min_X \sum_k \sum_d \sum_j \left(PD_{kdj} + ND_{kdj}\right) \tag{1}$$

$$subject\ to$$

$$c\min_k \leq \sum_d \sum_j x_{kdj} \leq c\max_k \quad \forall\, k \tag{2}$$

$$s\min_{dj} \leq \sum_k x_{kdj} \leq s\max_{dj} \quad \forall\, dj \tag{3}$$

$$h\min_{kd} \leq \sum_j x_{kdj} \leq h\max_{kd} \quad \forall\, k,d \tag{4}$$

$$b\min_{kdj} \leq x_{kdj} \leq b\max_{kdj} \quad \forall\, k,d,j \tag{5}$$

$$PD_{kdj} - ND_{kdj} = x_{kdj} - x_{0kdj} \quad \forall\, k,d,j \tag{6}$$

$$PD_{kdj} \geq 0, ND_{kdj} \geq 0, x_{kdj} \geq 0 \quad \forall\, k,d,j \tag{7}$$

The x_{kdj} represent suppressed employment data of 3-digit industries in county *k*. *dj* is the 3-digit industry code where *d* represents the 2-digit parent and *j* is the 3-digit level identifier (e.g., x_{k213} implies $d = 21$ and $j = 3$). x_{0kdj} is the target estimate related to x_{kdj}'s true value. x_{0kdj} is assumed as the product of the number of establishments of x_{kdj}, which is never suppressed, and the midpoints of the employee size classes in which the establishments are grouped (e.g., an x_{kdj} with only six establishments in size class 1 to 4 employees implies a midpoint estimate of 15 employees). PD_{kdj} and ND_{kdj} are non-negative variables that measure the positive or negative deviation, respectively, between x_{kdj} and x_{0kdj}. Objective function (1), in conjunction with constraint set (6), is equivalent to the problem of minimizing the sum of the absolute difference between the x_{kdj} and x_{0kdj}, as shown in *[33]*.

The first set of constraints ensures that the x_{kdj} sum is consistent with the provided total county employment boundaries $[cmin_k, cmax_k]$. The second set of constraints relates the sum of all suppressed dj industries at the county level with the known dj industry employment boundaries $[smin_{dj}, smax_{dj}]$ at the state level. In the third set of constraints, $hmin_{kd}$ ($hmax_{kd}$) is the minimum (maximum) 2-digit industry bound for industry dj in county k. Hence, this set of constraints guarantees that the sum of the 3-digit industries' employment is consistent with their respective 2-digit parent employment at the county level. The fourth set of constraints determines the bounds $[bmin_{kdj}\ bmax_{kdj}]$ given by the flags. Note that $cmin_k = cmax_k$, $smin_{dj} = smax_{dj}$, and $hmin_{kd} = hmax_{kd}$ if the quantity under consideration is known. Additionally, the bounds of the first three constraint sets are the employment information minus the known 3-digit employment. For example, if county k's total employment is 500 and the sum of all

unsuppressed 3-digit employment information is 450, then $cmin_k = cmax_k = 50$. For FAF regions without metropolitan subregions, the suppressed 3-digit information was imputed using national level data. The final set of imputed CBP data in California has been made public for other researchers and practitioners, and is available for download from the Cal-FRED web repository (freight.its.uci.edu/calfred) *[5]*.

For sub-county FAZs in California the imputed county level data was disaggregated using sub-county information such as land-use maps, ZIP code level employment information, and statewide farmland maps.

4 A STRUCTURAL COMMODITY GENERATION MODEL

4.1 Model Specification

Production and consumption of the commodity groups in Table 1 were modeled using FAF data for the U.S. Each commodity group is associated with production and consumption dependent variables, which are all estimated simultaneously, considering direct correlation and causal effects defined in a confirmatory manner using paths. The sample size is limited since there are 123 FAF regions in the U.S. and each zone is a sample. The small sample size means that we have to be very careful in specifying the model to avoid under-identification. In order to have an identified model, correlation is not assumed between every dependent variable and error term. Instead, those commodities with higher correlations are selected based on factor analysis when specifying the structural model. The model includes 12 of the 14 commodity production variables and 5 of the 14 commodity consumption variables, while the other commodities are estimated independently. The general formulation of the model is shown in Equation (8).

$$\begin{cases} P_i^m = \sum_k \beta X_{ik} + \gamma P_i^n + \delta C_i^n + e_{mi} \\ C_i^m = \sum_k \beta X_{ik} + \gamma P_i^n + \delta C_i^n + e'_{mi} \\ \mathrm{cov}(X_k, X_l), \mathrm{cov}(e_m, e_n), \mathrm{cov}(e_m, e'_n), \mathrm{cov}(e'_m, e'_n), \mathrm{cov}(P_i^n, e_m), \mathrm{cov}(C_i^n, e'_m) \epsilon\, R \end{cases} \quad (8)$$

P_i^m and C_i^m are production and consumption of commodity group m in zone i. X_{ik} are exogenous socioeconomic or industrial related explanatory variables and β, γ and δ are calibration parameters. e_{mi} and e'_{mi} are unobserved error terms in production and consumption equations. The model does not assume independence between explanatory variables (indicators) and error terms.

The model exhibits high skewness (3.5) and kurtosis (15.7), evidence of data with non-normal distributions. Kline *[34]* discussed different methods to deal with non-normal data in SEM. The asymptotically distribution free (ADF) method is the most popular method to deal with non-normality if the sample size is large (which this case is not). Maximum likelihood (ML) estimation is robust to non-normality, but the standard error is underestimated in the absence of the normality assumption. Nevitt and Hancock *[35]* recommended bootstrapping with more than 250 bootstrap samples for estimating the test statistic, the p values, and the standard errors under non-normal data conditions with original sample sizes greater than 100. In this study, 300 bootstrap samples resulted in meaningful analysis. All generated samples were usable.

Since SEM is a confirmatory approach, the covariance assumptions in the structure between endogenous and error residuals can be tested to find statistically significant relations.

For example, significant covariance was identified between the residuals of the consumption functions for commodity groups 3 (petroleum) and 4 (fuel and oil products). This could mean that there is some latent variable affecting consumption of both commodity groups 3 and 4.

A path diagram is a convenient method to represent SEM models. Figure 2 shows a path diagram of the proposed model drawn in SPSS AMOS. Straight single-end arrows show causality effects and double-end curved arrows show covariance effects *[34]*. The graph is color coded. Green arrows correspond to residuals. Blue arrows are secondary effects of consumption and production of different commodity groups on each other. Black arrows are causal effects of explanatory variables in production and consumption functions. Pink arrows correspond to correlation of exogenous variables. For simplicity of presentation, not all pink arrows are shown in this figure. A set of socioeconomic and industry related variables are chosen to estimate this model, with the estimated coefficients shown in Table 2, which is separated into production (2A) and consumption (2B) variables for presentation.

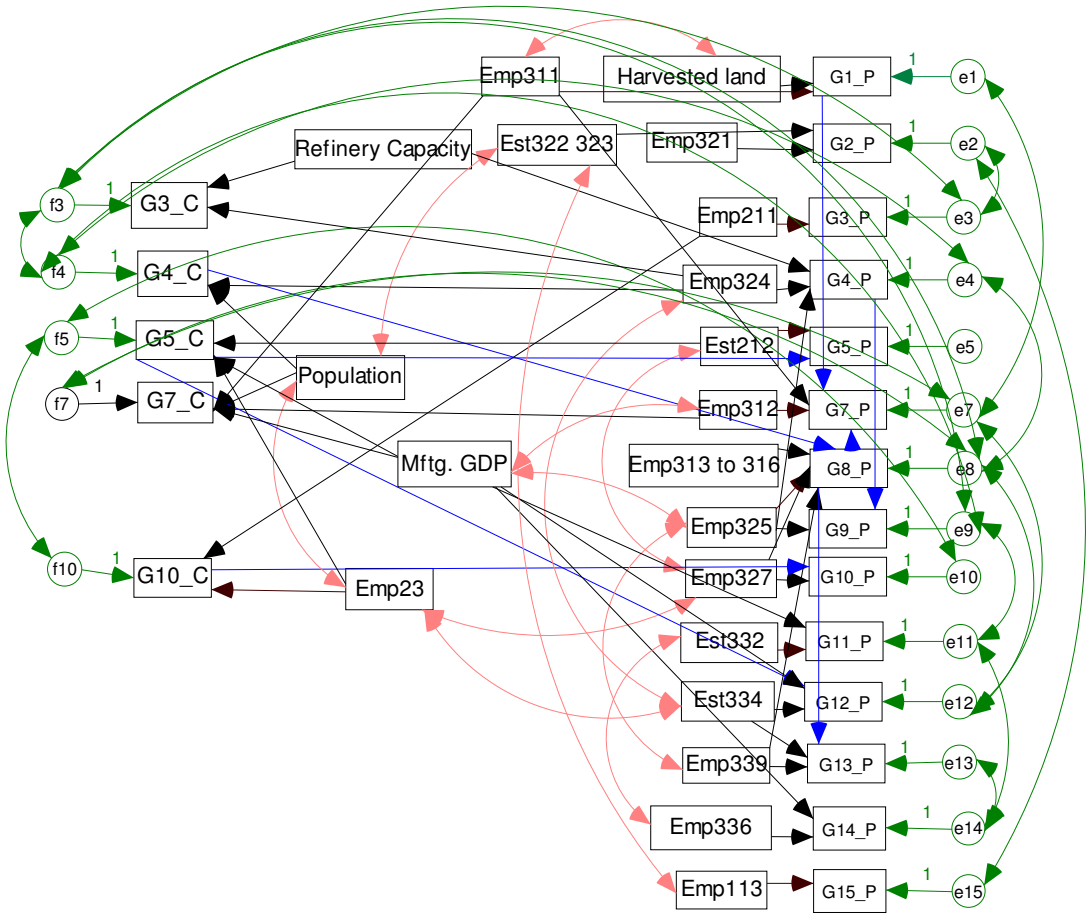


FIGURE 2. Path diagram for Structural Commodity Generation Model

TABLE 2A. Structural Commodity Generation Model (Production)

Dependent Variables	Independent Variables	Coefficient	Standardized Coefficient	Critical Ratio[†]
G1_P	Emp311(Food Mftg. Emp.)	0.434	0.151	2.95
	Harvested Land (acreage)	0.006	0.855	6.00
G2_P	Emp321 (Wood Products Mftg. Emp.)	0.794	0.739	8.82
	Emp322_323 (Paper Mftg., Printing & Related Support Activities Emp.)	4.894	0.3	4.490
G3_P	Emp211 (Oil & Gas Extraction Emp.)	4.63	0.783	3.481
G4_P	Emp324 (Petroleum & Coal Mftg. Emp.)	23.646	0.845	7.305
	Refinery Capacity (Barrels per Day)	0.006	0.061	0.545
G5_P	Est212 (No. Mining Establishments)	44.194	0.136	2.692
	G5_C	0.907	0.855	16.796
§G6_P	Emp212 (Mining Emp.)	8.24	0.536	7.005
	Est213 (Support Activities for Mining Emp.)	34.77	0.171	2.240
G7_P	Emp312 (Beverage & Tobacco Prod. Mftg. Emp.)	0.861	0.278	3.588
	Emp311(Food Mftg. Emp.)	0.194	0.368	4.042
	G1_P	0.056	0.307	4.000
	G8_P	0.337	0.39	3.210
G8_P	Emp325 (Chemical Mftg. Emp.)	0.142	0.155	1.560
	Emp313 to 316 (Textile, Apparel & Leather Mftg. Emp.)	0.179	0.26	3.729
	Emp339 (Misc. Mftg. Emp.)	0.221	0.189	2.125
	Emp327 (Nonmetallic Mineral Prod. Mftg. Emp.)	0.678	0.264	4.431
	G4_C	0.12	0.538	3.750
G9_P	Emp325 (Chemical Mftg. Emp.)	0.041	0.031	0.258
	G4_P	0.235	0.821	5.341
G10_P	Emp327 (Nonmetallic Mineral Prod. Mftg. Emp.)	0.179	0.049	1.32
	G10_C	0.927	0.956	23.77
G11_P	Est332 (No. Fabricated Metal Prod. Mftg. Establishments)	5.313	0.357	2.46
	Manufacturing GDP (Millions USD)	0.193	0.419	2.01
G12_P	Est334 (No. Computer & Electronic Prod. Mftg. Establishments)	42.085	0.685	7.46
	G5_C	0.163	0.277	7.09
	Manufacturing GDP (Millions USD)	0.091	0.11	1.90
G13_P	Emp339 (Misc. Mftg. Emp.)	0.039	0.272	1.39

	G8_P	0.059	0.492	3.93
	Est334 (No. Computer & Electronic Prod. Mftg. Establishments)	1.464	0.298	1.38
G14_P	Emp336 (Transportation Equp. Mftg Emp.)	0.082	0.545	2.41
	Manufacturing GDP (Millions USD)	0.062	0.29	2.82
G15_P	Emp113(Forestry & Logging Mftg. Emp.)	6.742	0.802	3.58

TABLE 2B. Structural Commodity Generation Model (Consumption)

Dependent Variables	Independent Variables	Coefficient	Standardized Coefficient	Critical Ratio†
§G1_C	Emp311 (Food Mftg. Emp.)	0.775	0.406	7.042
	Sold live stock (KTons)	0.012	0.554	9.606
§G2_C	Emp322 (Paper Mftg., Printing Emp.)	0.635	0.423	9.781
	Emp337 (Furniture Mftg. Emp.)	0.051	0.045	0.980
	Population	0.001	0.559	12.246
G3_C	Emp324 (Petroleum & Coal Mftg. Empl.)	3.086	0.272	3.040
	Refinery Capacity (Barrels per Day)	0.029	0.724	5.800
G4_C	Emp324 (Petroleum & Coal Mftg. Emp.)	21.444	0.857	4.886
	Population	0.003	0.179	3.000
G5_C	Est212 (No. Mining Establishments)	131.365	0.428	5.564
	Emp23 (Construction Emp.)	0.155	0.428	2.422
	Manufacturing GDP (Millions USD)	0.295	0.211	2.418
§G6_C	Coal power plants consumption(tons)	0.001	0.869	19.359
	Est213 (Support Activities for Mining	7.367	0.084	1.877
G7_C	Population	0.002	0.7	10.000
	Emp311 (Food Mftg. Emp.)	0.2	0.46	7.692
	Manufacturing GDP (Millions USD)	0.074	0.19	2.741
	Emp312 (Beverage & Tobacco Prod. Mftg. Emp.)	0.135	0.053	0.692
§G8_C	Population	0.002	0.513	9.119
	Emp325 (No. of Establishments Chemical Mftg.)	34.114	0.481	8.551
§G9_C	Emp325 (No. of Establishments Chemical Mftg.)	53.452	0.768	13.258
G10_C	Emp23 (Construction Emp.)	0.17	0.831	9.444
	Emp211 (Oil & Gas Extraction Emp.)	0.516	0.163	2.123
§G11_C	Emp332 (Fabricated Metal product Mftg. Emp.)	0.276	0.557	7.025

	Emp336 (Transportation Equp. Mftg Emp.)	0.086	0.199	3.965
	Manufacturing GDP (Millions USD)	0.110	0.233	3.034
§G12_C	Total Emp.	0.009	0.696	10.012
	Manufacturing GDP (Millions USD)	0.217	0.266	3.823
§G13_C	Population	0.0003	0.942	31.026
§G14_C	Manufacturing GDP (Millions USD)	0.043	0.287	3.895
	Emp336 (Transportation Equp. Mftg Emp.)	0.085	0.611	8.300
§G15_C	Emp321 (Wood Products Mftg. Emp.)	0.163	0.111	1.487
	Emp113 (Forestry and Logging Emp.)	6.465	0.758	10.130

§Estimated independently but included for completeness
†Critical Ratio = coefficient divided by bootstrap standard error

The first column is the dependent variable, the second column is the set of corresponding explanatory variables for each dependent variable, and the last three columns represent the coefficient value, standardized coefficient, and critical ratio. Most of the regression and covariance estimates are significant at the 0.05 level. Coefficients with lower significance are not discarded because this is the best result given available data, given the structural design that relates the dependent variables together.

4.2 Model Fitness Evaluation

There are different fitness measures in the SEM literature, as summarized by Hooper *et al. [36]*. Bollen and Long *[37]* suggested guidelines for presenting fitness indices of structural models. Table 3 summarizes the fitness indices of independent, hypothesized and saturated models.

TABLE 3. Model Fit Indices

Model Fit Index	Independent Model	Hypothesized Model	Saturated model
Sample size	119	119	119
Chi Square	8648	852	0.000
Degrees of Freedom (d.f.)	741	395	0
Goodness-of-fit index (GFI)	0.094	0.714	1.000
Normed fit index (NFI)	0.309	0.099	-
Akaike information criterion (AIC)	0.000	0.945	1.000
Incremental fit index (IFI)	0.000	0.942	1.000
Comparative fit index (CFI)	0.000	0.901	1.000
Expected Cross-Validation Index (ECVI)	70.937	13.091	12.559

The Chi-Square test can fail because the data is not multivariate normal, even though the model itself is properly specified. Since the data is indeed non-normal in this case, the Chi-Square test is disregarded. McIntosh *[38]* recommended that other fitness measures and predictive ability of the model should be considered in these cases. He suggested that "Merely setting for close fit could hinder the advancement of knowledge in a given substantive field, since there is little impetus to seek out and resolve the reasons why exact fit was not attained."

The Independent Model assumes all relationships among measured variables are 0 – it is the assumption of having no structure in place and serves as a baseline for comparison. The Saturated Model would perfectly reproduce all of the variances, covariances, and means; it has the best fit possible with zero degrees of freedom. The gap between the Independent and Saturated models signifies the degree of correlations present in the data that is not captured by the Independent Model. The proposed Hypothesized Model falls in between the two extremes; the closer it is to the Saturated Model and further from the Individual, the more the structural design is able to accommodate all the structural relationships present in the data. Clearly, the proposed model outperforms independent linear models in all measures shown in Table 3.

4.3 Analysis

Having shown that the structural model has a better fit, a few examples are discussed in this section to investigate the advantages of a structural commodity generation model for policy analysis.

The total effect (sum of direct and indirect effects) of each explanatory variable on production of each commodity group is presented in Table 4. The variables are in their original units. Since only five commodity consumption variables are included in the structural model, the total effects of those explanatory variables are similar to the coefficients and therefore left out of this table for brevity (readers are referred to Table 2B for essentially the same values). Effects of different variables on generation of freight in the entire system can be compared. For example, adding one employee in food manufacturing (Emp311) in a zone will generate about 434 tons of agricultural products (G1) and 218 tons of food and beverages products (G7) in a year.

TABLE 4. Total effects in structural production model

Variables	G1_P	G2_P	G3_P	G4_P	G5_P	G7_P	G8_P	G9_P	G10_P	G11_P	G12_P	G13_P	G14_P	G15_P
Mftg. GDP	-	-	-	-	0.267	-	-	-	-	0.194	0.138	-	0.062	-
Harvested land	0.006	-	-	-	-	-	-	-	-	-	-	-	-	-
Refinery Capacity	-	-	-	0.007	-	-	-	0.002	-	-	-	-	-	-
Population	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Est322 323	-	4.897	-	-	-	-	-	-	-	-	-	-	-	-
Est334	-	-	-	-	-	-	-	-	-	-	42.10	1.46	-	-
Est212	-	-	-	-	163.3	-	-	-	-	-	21.46	-	-	-
Est332	-	-	-	-	-	-	-	-	-	5.268	-	-	-	-
Emp113	-	-	-	-	-	-	-	-	-	-	-	-	-	6.746
Emp313 to 316	-	-	-	-	-	0.06	0.179	-	-	-	-	0.011	-	-
Emp336	-	-	-	-	-	-	-	-	-	-	-	-	0.082	-
Emp312	-	-	-	-	-	0.860	-	-	-	-	-	-	-	-
Emp23	-	-	-	-	0.141	-	-	-	0.158	-	0.025	-	-	-
Emp339	-	-	-	-	-	0.075	0.221	-	-	-	-	0.052	-	-
Emp211	-	-	4.632	-	-	-	-	-	0.478	-	-	-	-	-

Emp325	-	-	-	-	-	0.047	0.141	0.037	-	-	-	0.008	-	-
Emp327	-	-	-	-	-	0.229	0.679	-	0.179	-	-	0.04	-	-
Emp311	0.434	-	-	-	-	0.218	-	-	-	-	-	-	-	-
Emp324	-	-	-	23.31	-	0.869	2.576	5.483	-	-	-	0.153	-	-
Emp321	-	0.792	-	-	-	-	-	-	-	-	-	-	-	-

4.3.1 Indirect Effects

Table 4 can be used to analyze indirect effects. For example, as shown in Table 2, the factors for P_i^7 (food and beverage production) are: employment in food manufacturing (Emp311), employment in beverage and tobacco product manufacturing (Emp312), production of commodity group 1, and production of commodity group 8. It means production of commodity group 7 relates to production of two other commodity groups in each region. This is compared with an independent model estimated from similar data.

Structural model: $P_i^7 = 0.861Emp_{312} + 0.194Emp_{311} + 0.056P_i^1 + 0.337P_i^8$

Standardize structural model $P_i^7 = 0.278Emp_{312} + 0.368Emp_{311} + 0.307P_i^1 + 0.390P_i^8$

Independent model with standardized coefficients: $P_i^7 = 0.249Emp_{312} + 0.743Emp_{311}$

In this example, the effect of Emp312 on P_i^7 is very similar in both the structural and independent models. However, the effect of Emp311 on the independent model is 2.02 times greater than the structural model. The better fitting structural model suggests that Emp311 should actually be divided between its direct effect on P_i^7 and an indirect effect from its effect on P_i^1, which in turn has an effect on P_i^7.

4.3.2 Supply Chain Elasticity

The structural model is able to capture aggregate industry-level supply chain interactions so that elasticities can be measured. For example, Figure 2 shows that $\text{Population} \rightarrow \mathrm{C_i^4} \rightarrow \mathrm{P_i^8} \rightarrow \mathrm{P_i^7}$. Production of food products (G7) is affected by production of manufactured goods (G8), which are in turn affected by consumption of fuel and oil products (G4), which is affected by population. An increase in population by 10 results in 30 additional tons of fuel and oil consumption per year, which leads to 3.6 more tons of manufactured goods produced, resulting in 1.2 more tons of food, beverage, and tobacco products produced. This model framework gives a better understanding of supply chain elasticities between explanatory variables of different commodity groups which was largely ignored in previous freight transportation studies.

4.3.3 Intrazonal Freight

The model can capture the relationship between production and consumption of low value commodities that are typically not transported long distances. One example is commodity group 10, nonmetallic mineral products. This group is mainly composed of ready-mix concrete, which is a relatively cheap product and cannot be transported long distances due to its physical properties. Production of this product is highly driven by demand for it. In Figure 2, construction employment is the main driver for $\mathrm{C_i^{10}}$, which in turn is a factor for production of the commodity group: $Emp23 \rightarrow C_i^{10} \rightarrow P_i^{10}$. Whereas an independent model would likely have a difficult fit for P_i^{10}, this model indicates that one additional employee in the construction industry in a region

could lead to 170 tons of nonmetallic mineral products consumed in that same region, which results in 927 tons of that product being produced in the same region.

5 VALIDATION OF GEOGRAPHIC SCALABILITY WITH CALIFORNIA FREIGHT ANALYSIS ZONES

The calibrated model was applied to the California FAZs using 2007 imputed data to investigate the geographic scalability of the model. The forecast productions and consumptions of each commodity group based on the structural model using the imputed data at the FAZ level were then aggregated up to the FAF level for comparison. Two commodity groups are shown in Figure 3 to illustrate the distribution of production forecast across FAZs in California.

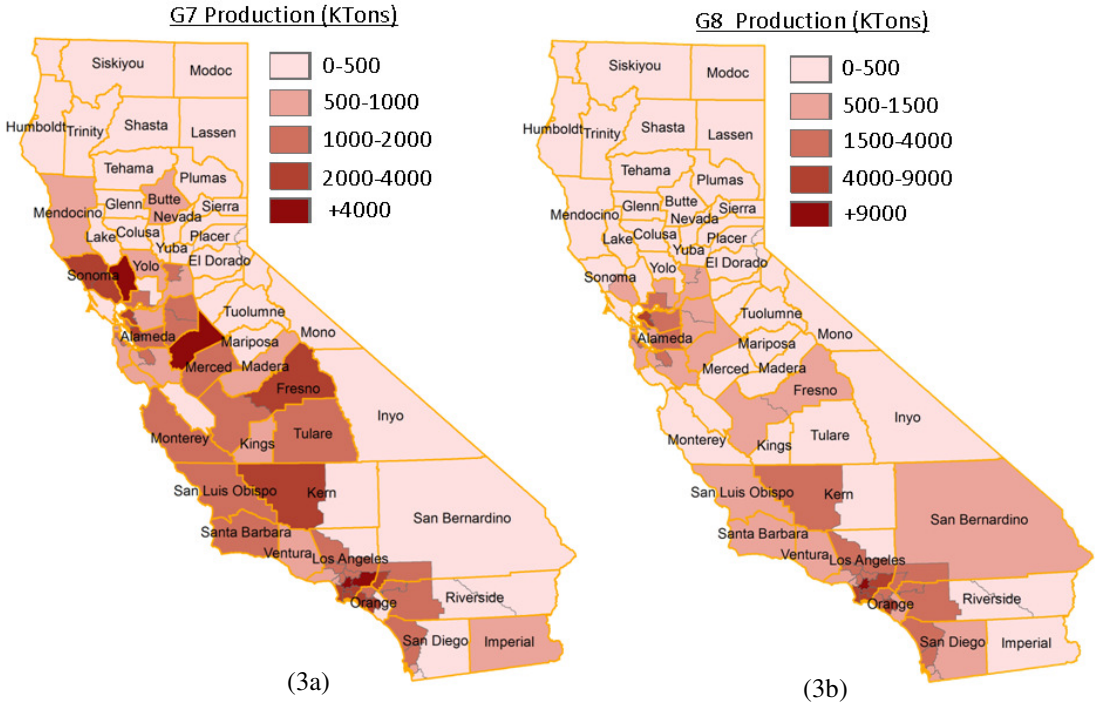
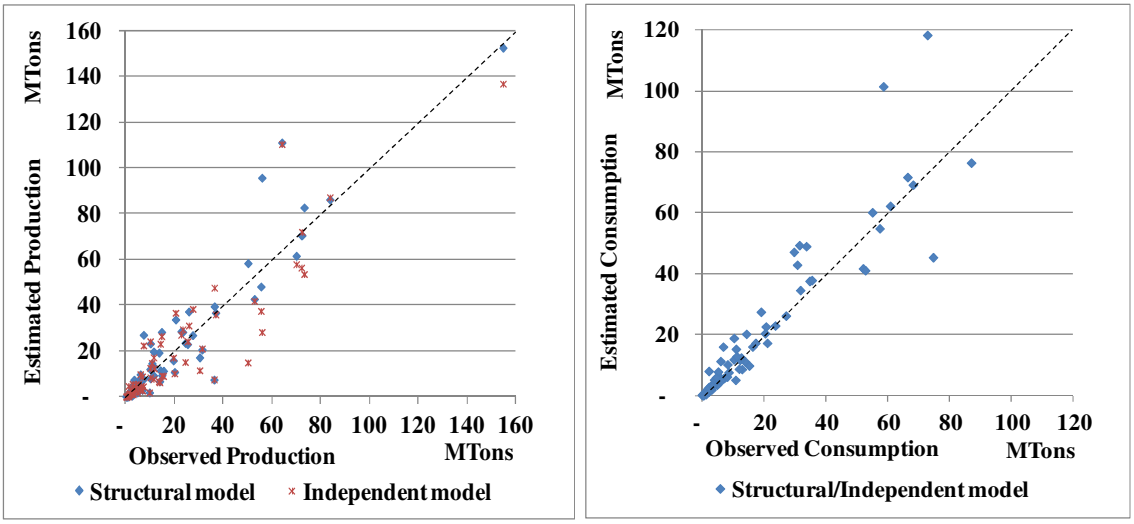


FIGURE 3. Comparison of Production of commodity group 7 (3a) and 8 (3b) in California FAZs.

The forecast FAZ production and consumption values were then aggregated back up to the FAF level for validation. Figure 4a compares the FAF-level observed and estimated production of all commodity groups, except six (14 groups x 5 x 2 data points = 140 observations). Commodity group 6 (coal and metallic ores) is excluded because there is no production of coal or metallic ores in California and consumption is limited to three small coal power plants and a few firms. Since the consumption of all commodities was not included in the structural model due to limited sample size, the production model is a mixed structural/independent model.

The mean absolute percentage error (MAPE) for the California FAF-level production model and aggregated FAZ-level are respectively 38.4% and 38.9%. In other words there is only a 0.5 percentage point loss of accuracy between these two models. On the other hand, the MAPE for the aggregated independent FAZ production model is 44.8%. In fact, the present error appears to be primarily from estimating with national level data and applying it to California, not

from scaling from FAF regions to FAZs. The total production in California predicted by the structural model for the FAZs differs from the total observed production by 2%. As for the consumption model, the MAPE also remains unchanged from the FAF-level to FAZ-level at 33%. The total consumption in California predicted by the model differs by 7.9% from the total observed consumption. The results demonstrate the improved robustness of the structural model and the geographic scalability of the estimated parameters and hypothesized structure.



(4a) (4b)

FIGURE 4. Geographic scalability. Structural production model (4a) versus independent production model (4b).

6 CONCLUSION

This is the first study to propose and geographically validate a structural commodity generation model using only public data. SEM can capture inter-dependencies between different commodity groups and production and consumption, effectively inferring the unobserved supply chain, land use, and intrazonal supply-demand relationships without sacrificing geographical scalability.

Several interesting empirical findings were made with the structural model. Independent models may overemphasize the direct effects of certain explanatory variables because of the inability to capture indirect effects. Several supply chain interactions can be identified by the model, such as the relationship between fuel and oil consumption, production of manufactured goods, and production of food, beverage, and tobacco products. The structural model is also able to explain intrazonal freight generation, such as ready-mix concrete.

The results of this study can help to identify major data gaps and the contribution of each variable in the entire model. The user can prioritize the data needs to improve overall results based on the covariance matrix and total effect of each variable. The covariance matrix shows where the relationship between dependent and independent variable is weak, which may result in poor estimation and new variables are then required to fill the data gap. Also, improving the quality of variables with highest total effect will improve the model fitness.

Given the aggregate nature of public data, the accuracy of the results may be affected by how poorly some commodity groups were captured in FAF. For example, farm based products,

fisheries, logging, construction and crude petroleum are not included in CFS. FAF used other data sources to estimate the flow of these commodities. On the other hand, manufactured products, electronics, and food and beverages are covered more in the CFS sample, so we were able to get better results for these groups. The results of this study serve as a benchmark so that future replication studies with improved data or improved structural designs can build from this foundation.

Future studies should consider further specifications of commodity generation, distribution, or integrated land use models based on a SEM structure. Studies of this nature on both commodity and vehicle based freight transportation in an urban setting would be of interest to researchers and practitioners alike. This model framework can be readily integrated with disaggregate models (agent-based demand models, firm-based mode/shipment choice models) or assignment and simulation models (transshipment, scheduling, and queueing optimization or simulation).

7 ACKNOWLEDGMENTS

The research reported in this paper was supported by the California Department of Transportation. The authors gratefully acknowledge the assistance provided by Doug MacIvor, Chad Baker, Kalin Pacheco and Diane Jacobs of the California Department of Transportation. Helpful comments from four anonymous reviewers are appreciated. The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California. This paper does not constitute a standard, specification, or regulation.

8 REFERENCES

[1] *MAP-21: A Summary of Highway Provisions*. FHWA. http://www.fhwa.dot.gov/map21/summaryinfo.cfm. Accessed July 27, 2012.

[2] *NCHRP Report 606: Forecasting Statewide Freight Toolkits.* Transportation Research Board of the National Academies, Washington, D.C., 2008.

[3] Chow, J.Y.J., C.H. Yang, and A.C. Regan. State-of-the art of freight forecast modeling: lessons learned and the road ahead. *Transportation*, Vol. 37, No. 6, 2010, pp. 1011-1030.

[4] Tavasszy, L.A., K. Rijgrok, and I. Davydenko. Incorporating logistics in freight transport demand models: state-of-the-art and research opportunities. *Transport Reviews* Vol. 32, No. 2, 2012, pp. 203-219.

[5] Tok, A.Y.C., M. Zhao, J.Y. J. Chow, S.G. Ritchie, and D.I. Arkhipov. Online data repository for statewide freight planning and analysis. In *Transportation Research Record: Journal of the Transportation Research Board,* No. 2246, 2011, pp. 121-129.

[6] Kockelman, K.M., L. Jin, Y. Zhao, and N. Ruiz-Jur. Tracking land use, transport, and industrial production using random-utility-based multiregional input-output models: applications for Texas trade. *Journal of Transport Geography,* Vol. 13, No. 3, 2005, pp. 275-286.

[7] Zhong, M., J.D. Hunt, and J.E. Abraham. Design and development of a statewide land use transport model for Alberta. *Journal of Transportation Systems Engineering and Information Technology*, Vol. 7, No. 1, 2007, pp. 79-91.

[8] Park, J.Y., J.K. Cho, P. Gordon, J.E. Moore, H.W. Richardson, and S.S. Yoon. Adding a freight network to a national interstate input-output model: a TransNIEMO application for California. *Journal of Transport Geography*, Vol. 19, No. 6, 2011, pp. 1410-1422.

[9] Donnelly, R. *A hybrid microsimulation model of urban freight transport demand*. Ph.D. dissertation. University of Melbourne, Melbourne, 2009.

[10] Liedtke, G. Principles of micro-behavior commodity transport modeling. *Transportation Research Part E*, Vol. 45, No. 5, 2009, pp. 795-809.

[11] Roorda, M.J., R. Cavalcante, S. McCabe, and H. Kwan. A conceptual framework for agent-based modelling of logistics services. *Transportation Research Part E*, Vol. 46, No. 1, 2010, pp. 18-31.

[12] Holmgren, J., P. Davidsson, J.A. Persson, and L. Ramstedt. TAPAS: a multi-agent-based model for simulation of transport chains. *Simulation Modelling Practice and Theory*, Vol. 23, 2012, pp. 1-18.

[13] Giuliano, G., P. Gordon, Q. Pan, J.Y. Park, and L.L. Wang. Estimating freight flows for metropolitan area highway networks using secondary data sources. *Networks and Spatial Economics,* Vol. 10, No. 1, 2010, pp. 73-91.

[14] Anderson, M.D., G.A. Harris, and K. Harrison. Using aggregated federal data to model freight in a medium-sized community. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2174, 2010, pp. 39-43.

[15] Novak, D.C., C. Hodgdon , F. Guo, and L. Aultman-Hall. Nationwide freight generation models: a spatial regression approach. *Networks and Spatial Economics*, Vol. 11, No. 1, 2011, pp. 23-41.

[16] Cambridge Systematics. *Development of a computerized method to subdivide the FAF2 regional commodity OD data to county level OD data*. FHWA, 2009.

[17] Holguín-Veras, J., M. Jaller, L. Destro, X.J. Ban, C. Lawson, and H.S. Levinson. Freight generation, freight trip generation, and perils of using constant trip rates. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2224, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 68-81.

[18] You, S.I. *Methodology for Tour-based Truck Demand Modeling using Clean Truck at Southern California Ports*. Ph.D. dissertation. Department of Civil and Environmental Engineering, University of California, Irvine, 2012.

[19] Hunt, J.D., K.J. Stefan, and A.T. Brownlee. Establishment-based survey of urban commercial vehicle movements in Alberta, Canada. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1957, 2006, pp. 75-83.

[20] *Transportation Commodity Flow Survey 2007*. EC07TCF-US, Bureau of Transportation Statistics and U.S. Census Bureau, 2010.

[21] Aultman-Hall, L., B. Johnson, and B. Aldridge. Assessing potential for modal substitution from statewide freight commodity flow data. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1719, 2000, pp. 10-16.

[22] Ben-Akiva, M., and G. de Jong. The Aggregate-Disaggregate-Aggregate Freight Model System. In *Recent Developments in Transport Modelling: Lessons for the Freight Sector*,

(Ben-Akiva, M., Meersman, H., and Van de Voorde, E., eds.), Emerald Group Publishing Ltd., 2008, pp. 117-126.

[23] *NCHRP Synthesis 298: Truck Trip Generation Data*. Transportation Research of the National Academies. Washington, D.C., 2001.

[24] Cambridge Systematics, *Quick Response Freight Manual*, Report DOT-T-97-10, U.S. Department of Transportation and U.S. Environmental Protection Agency, Washington, D.C., 1997.

[25] Southworth, F., 2003. Freight Transportation Planning: Models and Methods. In *Transportation Systems Planning: Methods and Applications*. (Goulias, K.G., ed.) CRC Press, 29p.

[26] Bastida, C., and J. Holguín-Veras. Freight generation models: comparative analysis of regression models and multiple classification analysis. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2097, 2009, pp. 51-61.

[27] Chun, Y., H. Kim, and C. Kim, 2012. Modeling interregional commodity flows with incorporating network autocorrelation in spatial interaction models: an application of the US interstate commodity flows. *Computers, Environment and Urban Systems*, in press, http://dx.doi.org/10.1016/j.compenvurbsys.2012.04.002

[28] Golob, T.F. Structural equation modeling for travel behavior research. *Transportation Research Part B*, Vol. 37, No. 1, 2003, pp. 1-25.

[29] Jonnavithula, S.S. *Development of structural equations models of statewide freight flows*. Master's thesis. Department of Civil and Environmental Engineering, University of South Florida, Tampa, 2004

[30] *County Business Patterns 2007*. U.S. Census Bureau. http://www2.census.gov/econ2007/CB/sector00/CB0700A2.zip. Accessed Feb. 23, 2012.

[31] *Census of Agriculture 2007*. U.S. Department of Agriculture http://www.agcensus.usda.gov/Publications/2007/index.php. Accessed May 21, 2012

[32] *Regional Economic Accounts*. Bureau of Economic Analysis, U.S. Department of Commerce. http://www.bea.gov/regional/index.htm. Accessed May 29, 2012.

[33] Zhang, S., and J.M. Guldmann. Estimating suppressed data in regional economic databases: A goal-programming approach. *European Journal of Operational Research*, Vol. 192, No. 2, 2009, pp. 521-537.

[34] Kline, R.B. *Principles and Practice of Structural Equation Modeling.* 2nd ed., The Guilford Press, New York, 2005.

[35] Nevitt, J., and G.R. Hancock . Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, Vol. 8, No. 3, 2001, pp. 353–377.

[36] Hooper, D., J. Coughlan, J., and M. Mullen, 2008. Structural equation modelling: guidelines for determining model fit. *Electronic Journal of Business Research Methods*, Vol. 6, No. 1, 2008, pp. 53-60.

[37] Bollen, K. A., and J. S. Long. Introduction. In *Testing Structural Equation Models* (K. A. Bollen and J. S. Long, eds.), Sage Publications, Newbury Park, California, 1993, pp. 1–15.

[38] McIntosh, C.N., 2007. Rethinking Fit Assessment In Structural Equation Modeling: A Commentary And Elaboration on Barrett (2007) Vol. 42, No.5,2007, pp. 859‑867.