Freight Model Validation Techniques

Abstract

Several reviews of validation techniques for statewide passenger and freight models have been published over the past several years. In this paper I synthesize these studies and highlight validation techniques that require observed data related to truck counts, classification, and travel patterns.

Introduction

Statewide freight models provide freight performance and demand information necessary to make policy decisions for future years. It is important for freight forecasts provided by state wide models to be accurate in their predictions as they can affect policy decisions regarding pavement and safety management, project prioritization, modal diversion, port and terminal management, and time-of-day shift policies. The process of determining the level of accuracy of freight models in predicating future freight flows in referred to as model validation. A more formal definition of validation by the National Cooperative Highway Research Program (NCHRP) is "the process of comparing model outputs against data to determine how well the model simulates *aggregate measurements* of behavior " (1). Further, since the final output of statewide freight models is often link level truck counts, validation can be defined as "an analysis of a travel demand model based on *traffic count* and other information" (4).

The NCHRP has published several reports which summarize statewide freight model research efforts including NCHRP Report 606 Forecasting Statewide Freight Toolkit (1) and NCHRP Project 836-B Task 91 Validation and Sensitivity Considerations for Statewide Models (2). These reports discuss freight demand forecasting models used by various states' as well as recommend frameworks for model calibration and validation. In addition, the Federal Highway Administration (FHWA) outlines freight model validation and calibration techniques in their publication, the Quick Response Freight Manual (3). While each of these reports highlights and suggests validation techniques, a review of statewide freight models makes it clear that validation of statewide freight models is not a necessary requirement. In fact, the reports state that "some components of freight models are typically not validated since the only data available was used to develop the model and no independent data are available for validation" (1). This is quite alarming considering the wide range of policy decisions that can be drawn from freight flow predictions. Thus, the purpose of this paper is to review freight model validation techniques outlined by the NCHRP and the FHWA and to summarize selected state

freight model validation efforts. Further, since highway freight transport accounts for the majority of freight flow in the U.S., in this paper, validation techniques which require observed data related to truck counts, classification, and travel patterns will be highlighted.

This paper is divided into five sections. First, data sources for validation are summarized. Second, model validation techniques including reasonableness checks and statistical methods are presented. Third, a brief discussion of validation techniques for various freight model types is provided. Fourth, the Florida Statewide Model validation approach is given as a case study. Finally, further questions and conclusions are offered in the fifth section.

Model Validation Data Sources

Statewide freight models are compiled from many data sources and likewise use a variety of data sources for validating their results. It is important to note that a data source can only be used for validation of a model if that source was not used in building/calibrating the model. For example, if a commodity flow survey was used as a basis for developing or calibrating a commodity based freight forecasting model, then that same survey should not be used to validate the model. While this may seem like an obvious requirement, freight data sources are so sparse and lacking that some states can only validate on the same sources that were used to build their models. One way to cope with this shortcoming will be presented in a later section (i.e. 'backcasting'). The data sources presented in this section represent existing national, state, and local sources that have been used by states in validation efforts. Additionally, states may also choose to collect additional data specific to their modeling efforts.

Truck surveys

Intercept surveys, or origin–destination surveys, provide data on trip patterns, commodity type, vehicle type, and trip lengths. Origin–destination surveys are critical for trip generation and distribution and, particularly, for understanding external trip movements. For freight purposes, travel surveys take the form of shipper and carrier surveys in which a statistically valid sample is collected. This can be expensive and difficult to conduct and is therefore not generally carried out by each state but instead conducted as part of the national Commodity Flow Survey (1). The Commodity Flow Survey, TRANSEARCH database, and the Freight Analysis Framework are several examples of national level shipper survey databases that have been used in validation (1).

Highway Performance Monitoring System

A common validation technique is to compare final model vehicle-miles-travelled (VMT) to observed VMT at the state level (2). The FHWA's Highway Performance Monitoring System (HPMS) is used to compare model VMT against estimates by functional class and area type.

State DOTs are required to include Annual Average Daily Traffic (AADT) Counts and mileage for roadways based on a statistical sample, for each urban area as part of their annual HPMS submittal.

Vehicle Inventory and Use Survey

Collected by the U.S. Census Bureau every five years, the Vehicle Inventory and Use Survey (VIUS) data can provide highly useful data for model validation, including major use, products carried, annual and lifetime miles, area of operations, and fuel usage characteristics (1). Most important for validation are total VMT and trip length data by vehicle type and commodity at the state and national levels (2). As with truck surveys, a major drawback is that the VIUS data represents a sample of all vehicles, approximately 2,000 vehicles per state (1).

Truck classification and weight data

Truck classification and weight data comes from Weigh-in-motion (WIM) sites from each state. The Vehicle Travel Information System (VTRIS) is designed to provide a standard format for presenting the outcome of vehicle weighing and classification efforts at truck weigh sites for vehicle classification counts and can be a valuable tool for validation. Vehicle classification counts are "one of the only sources to verify the reasonableness of traffic volumes based on the inclusion of commercial vehicles into the transportation planning models" (3). Counts are available for FHWAs 13 axle based classes by location and time of day. It should be noted that freight models which predict only freight demand will provide estimates less than total truck counts since truck counts include both freight and non-freight movements. WIM data can only go as far as to provide total truck counts since axle configuration and weight alone do not determine if a truck carries freight or has another commercial purpose.

Registration records

State vehicle registration records can be used for comparison of model fleet sizes to observed fleet sizes (3). Commercial/service use is not explicitly recorded by registration records but can be inferred from the vehicle make/model, weight class, and body-type as was done by Cambridge Systematics in a federal study (3). Care has to be taken in using vehicle registration records because trucks operating in a certain state may actually be registered in another state.

Tolled Facility Electronic Data

A more advanced form of data involves tracking trucks thru tolled facilities by collecting electronic toll tag data. Tolled facility data could provide origin-destination information to verify freight model estimates. This type of data is only available for a small portion of facilities and has only been used on a study-by-study basis (1).

Model Validation Techniques

The model validation data sources listed in the previous section can be used for direct statistical comparison and more broadly as sources for assessing the reasonableness of estimated results. For freight models with intermediate outputs, such as the four step model, validation should be carried out for each individual step, i.e. trip generation and trip distribution, as well as for the overall model outputs. In this section, types of reasonableness checks and common statistical tests are presented.

Reasonableness Comparisons

Reasonableness comparisons involve comparing the ranges and magnitudes of model parameters, rates, and intermediate and final outputs with those found in other reports, models, or regions. For example, the percent of VMT per household representing commercial vehicle travel should be within a reasonable range when comparing two regions with similar characteristics. In fact, the Quick Response Freight Manual provides ranges and suggests alternative sources such as relevant NCRHP reports for reasonableness ranges for many of the freight model parameters and rates (3). Reasonableness comparisons are specifically useful for trip generation rates and logit model mode choice parameters. However, because the geographic and economic setting for each state is so different, it is not recommended to validate the results with other statewide models even though they may have similar characteristics such as trip generation, trip length, or mode split. Instead, reasonableness checks to the Quick Response Freight Manual should be conducted (3).

Rather than comparing state –to- state or state- to -national model parameters, rates, and results, comparison of statewide to urban or regional model results can be carried out. For example, the Florida Statewide Model uses statistical comparisons between regional models within districts to further validate district level models and comparisons of screenline counts to compare district and statewide models (4). Discrepancies between model results at the region, district, and statewide levels may arise for average trip lengths since statewide models emphasize longer distance trips, i.e. freight based trips as opposed to commercial trips (2).

The final form of reasonableness comparisons is backcasting. Backcasting is the process of estimating freight flows for a year prior to the year on which a freight model was developed and calibrated. For instance, forecasts for year 2002 could be estimated from a model developed and calibrated on year 2007 data. Issues for this method include data format and consistency problems, for example, data available in 2007 and beyond may not be available for previous years. The UK Transport Model, while not specifically for freight modeling, relied on a backcasting approach for validation in which two separate backcasts were performed

spanning 10 and 25 years back from the model base year (5). The authors deemed the backcasting approach the "most valuable source of validation evidence" for their model (5).

Statistical Analyses

The most commonly reported statistics for freight models are terminal times by purpose, average trip length by purpose, volume-over-count ratios by facility type or functional class and screenline, absolute difference between volume and counts, R-squared values of volumes versus counts, and root mean square error (RMSE) by volume group, facility type, or functional class (6). Several states have published accuracy standards and benchmarks including California, Michigan, Oregon, and Tennessee (6). In this section, common statistical analyses found in freight model validation studies are briefly described.

Volume-over-count ratios compare the modeled volume to the observed count in a simple ratio. Table 2 shows a sample of acceptable and preferable thresholds for volume-over-count ratios used in the Florida Statewide model. The volumes in this table refer to total volumes, that is, passenger, commercial, and freight traffic volume. Typically thresholds vary by facility type and higher order facilities such as freeways have more stringent standards. Table 1 shows an example of percent error thresholds based on 'volume groups'. Only a desired percentage of links within each volume group need to meet the standard. These two tables refer to statewide passenger models results; however, the same accuracy standards can be used for integrated statewide models, such as the Florida Statewide Model which combines freight and passenger travel for trip assignment (4). Additionally, the QRFM sets accuracy targets for aggregate level VMT for three categories of commercial vehicles, one of which includes vehicles for goods movement. The threshold for travel related to the movement of goods is 1-7% of total VMT and 3.5% average VMT according to the QRFM (6).

Table 1 Percent Error by Volume Group, Ref: FSUTMS (6)

	Standards	
Statistic	**Acceptable**	**Preferable**
Percent Error: LT 10,000 volume (2L road)	50%	25%
Percent Error: 10,000-30,000 (4L road)	30%	20%
Percent Error: 30,000-50,000 (6L road)	25%	15%
Percent Error: 50,000-65,000 (4-6L freeway)	20%	10%
Percent Error: 65,000-75,000 (6L freeway)	15%	5%
Percent Error: GT 75,000 (8+L freeway)	10%	5%

Table 2 Volume-over-count Ratios and Percent Error, Ref: FSUTMS (6)

Statistic	Standards	
	Acceptable	Preferable
Freeway Volume-over-Count (FT1x, FT8x, FT9x)	+/- 7%	+/- 6%
Divided Arterial Volume-over-Count (FT2x)	+/- 15%	+/- 10%
Undivided Arterial Volume-over-Count (FT3x)	+/- 15%	+/- 10%
Collector Volume-over-Count (FT4x)	+/- 25%	+/- 20%
One way/Frontage Road Volume-over-Count (FT6x)	+/- 25%	+/- 20%
Freeway Peak Volume-over-Count	75% of links @ +/-20%; 50% of links @ +/-10%	
Major Arterial Peak Volume-over-Count	75% of links @ +/-30%; 50% of links @ +/-15%	
Assigned VMT-over-Count Areawide	+/-5%	+/-2%
Assigned VHT-over-Count Areawide	+/-5%	+/-2%
Assigned VMT-over-Count by FT/AT/NL	+/- 25%	+/- 15%
Assigned VHT-over-Count by FT/AT/NL	+/- 25%	+/- 15%

The QRFM sets additional standards for counts at screenline locations. Figure 1 shows the maximum desirable deviation for total screenline volumes. For low screenline traffic volumes, the deviation from the observed volume can be as high as 65%.

Root Mean Square (RMSE), shown in Equation 1, is used to compare observed versus estimated volumes for all links with counts. As with percent error and volume-over-count ratios, RMSE acceptable ranges vary based on the facility type and/or volume group. For example, RMSE should be below 5% for freeways and below 40 to 50% for local and minor arterials (3).

$$\%RMSE = \frac{(\sum_j (Model_j - Count_j)^2 / (NumberofCounts - 1))^{0.5} * 100}{(\sum_j Count_j / NumberofCounts)}$$

Equation (1)

The Coefficient of Determination, R-squared, is used to determine the ability of a model to predict traffic volumes. R-squared is used as a statistical measure for comparing region-wide observed traffic counts to estimated volumes. Additionally, R-squared values can be used as reasonableness checks for Trip Generation regression models.

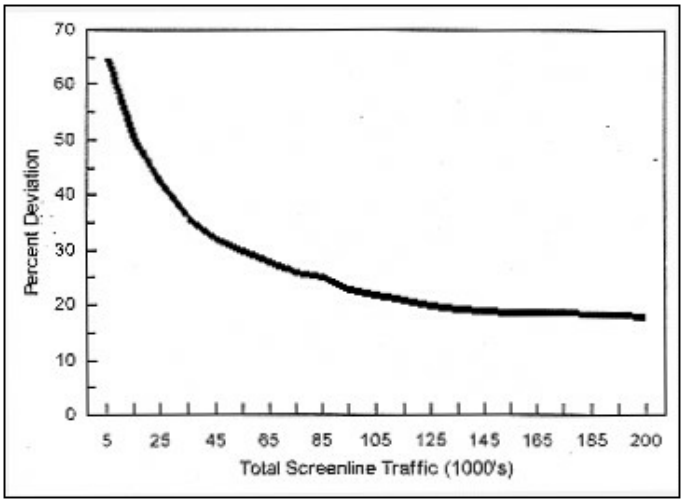

Figure 1 Maximum Desirable Deviation in Total Screenline Volumes, Ref: QRFM (3)

The GEH statistic calculates the difference between the observed and assigned traffic volumes results from final model outputs (see Equation 2). The GEH statistic is useful because one threshold can be set for all volume ranges, whereas for RMSE or percent error, as was seen previously, various thresholds were needed to account for the differing facility volumes. The Wisconsin Statewide Model is cited as using the GEH statistic as validation for the trip assignment step of the freight model (2).

$$GEH = \sqrt{\frac{(AssignedVolume - Count)^2}{((AssignedVolume + Count)/2)}}$$ Equation 2

RMSE, Volume-over-count ratios, percent error, and GEH statistics focus on the final outputs of freight models, the estimated traffic volumes. It is important also to validate intermediate steps such as trip distribution. Trip length frequency distribution is a useful measure for validation of trip distribution. The Coincidence ratio is a statistical comparison of frequency distributions for comparing observed and predicted trip length frequencies. The coincidence ratio measures the percent of area that "coincides" for two curves and ranges between zero, meaning two disjoint distributions, and one (3, 6). Figure 2 shows an example of 'good' and 'poor' coincidence ratios. For passenger statewide models the desired target for the coincidence ratio is between 65 and 70% (6). Whether or not this is acceptable for freight forecast models, is unclear.

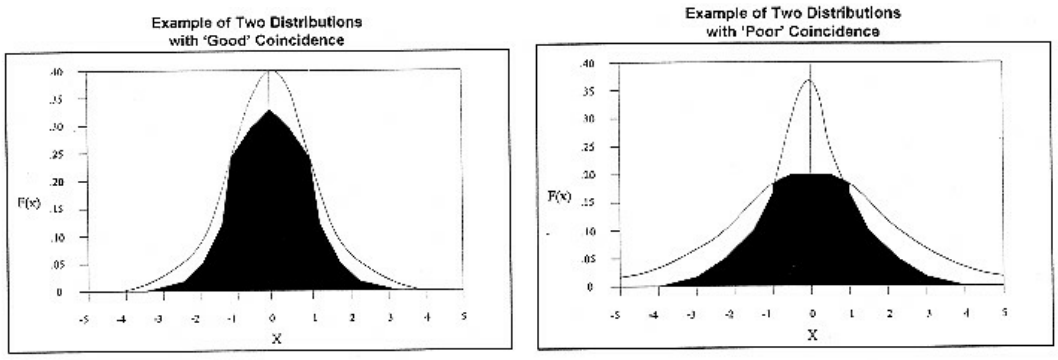

Figure 2 Coincidence Ratio Example, Ref: QRFM (3)

The statistical measures described in this section represent the most common measures found for statewide models and most were derived for passenger models final trip assignment validation. Whether or not the benchmark and threshold values for passenger travel are acceptable for freight forecast models is not addressed in the validation reports. Also, some states use metropolitan or urban model validation targets such as those found in the *NCHRP Report 255: Highway Traffic Data for Urbanized Area Project Planning and Designs* which might also not be appropriate for statewide freight models (1).

Validation Techniques by Freight Model Type

NCHRP Project 836-B Task 91 categorizes Freight models into four types: truck models, direct commodity table freight models, four step freight models, and economic activity models (2). In this section, the measures and estimates used for validation of each of the four model types are summarized.

The truck model does not distinguish between freight and non-freight trucks and therefore only needs to validate against total truck counts such as Annual Average Daily Traffic (AADT) volumes from state DOT databases. Other forms of validation include comparisons to previous commercial vehicle/truck surveys trip generation rates, trip length frequency distributions, and average trip lengths. Additionally, although not common, origins, destinations, and trip purposes can be validated through truck intercept surveys.

Direct Commodity Table Freight Models use directly acquired forecasts of commodity flows to forecast freight truck demand. Validation involves comparison of network assignment truck volumes which result from another commodity flow database to those resulting from the Direct Commodity Table Freight Model, e.g. if the Direct Commodity Table is developed from TRANSEARCH, than validation would use FAF2 to compare resulting truck volumes. Also,

comparison of payload factors (factors converting tons to truck loads) should be validated against independent sources such as VIUS.

Four step Freight models forecast multimodal flow of freight starting with explanatory variables for state/regional characteristics and results in forecasts of freight trucks and other non-highway modes. Trip Generation validation includes comparison of total truck trip productions and attractions per employee to national average rates, comparison of total truck trips by purpose to other models, regions, or agencies, and reasonableness of R-squared determination for generation regression models. Trip Distribution validation includes average trip length comparison to VIUS rates or trucker surveys, comparisons of trip length frequency distributions via the coincidence ratio, and comparison of estimated friction factors (from the gravity model) to other region's values. Mode Split validation includes comparing mode split coefficients from the logit mode choice model to other studies, and comparing the observed shares of freight flows to national databases. Trip assignment validation includes comparisons of VMT to HPMS VMT, vehicle classification counts at screenlines, and model fleet sizes to registration records. Four step models should be validated at each stage although validation of trip assignment results is the most important (3).

Lastly, economic activity models explain the interaction between the transportation system and economic activity and forecast freight trucks based on this interaction. In short, economic activity models use similar validation techniques as direct commodity models, mostly relying on validation at the assignment stage.

The Florida Intermodal Statewide Highway Freight Model

The Florida Intermodal Statewide Highway Freight Model is a four step commodity forecasting model and was selected as a case study because of the detailed documentation concerning its validation efforts (4). The validation process looked at outputs for all four steps in the model chain, rather than just the resulting assignments. Validation data consisted of truck counts by vehicle class from the 1999 AADT Report for Florida, the Truck Weight Study Data for the U.S., and TRANSEARCH. Additionally, the FAF loaded highway network was used to compare the percentage of freight trucks from the AADT data.

The validation approach followed a tiered methodology consisting of systemwide, districtwide, and corridor-level analysis. Each tier of validation used different techniques, comparisons, and accuracy thresholds. The systemwide validation process focused on statewide statistical evaluations at each stage of the four step modeling process. The districtwide validation focused on statistical comparisons to FDOT district freight models and district boundary counts.

The corridor level validation compared specific corridors comparisons between the statewide estimates and urban model estimates.

Conclusion

Of the validation techniques and data sources covered in this paper, much of the focus has been on comparisons of final model outputs to observed count data. Validation of truck types is not given as much weight as truck counts. Rather than simple comparisons of VMT by facility type, more detailed comparisons of VMT by truck type for each facility are important to validate. However, it is difficult to do this because truck type information is only available for a limited number of facilities such as weigh-in-motion stations located on major freeways. A solution to this problem is to implement traffic counting stations, i.e. loop detectors, with advanced detector technology that is capable of capturing vehicle types.

Further, in examining the thresholds and benchmarks set forth by the QRFM or the NCHRP synthesis reports, it is apparent that there is a level of inaccuracy in model outputs that is considered acceptable. In fact, depending on the facility volume, percent error between observed and estimated counts can be as high as 50% and as low as 10% (see Table 1). This begs the question; how sensitive are policy decisions to freight forecasts? The forecast may be unreliable but if the policies being analyzed don't really change within the confidence range then it doesn't really matter that the forecast is unreliable.

References

(1) Cambridge Systematics, Inc., Global Insight, Cohen, H., Horowitz, A., and Pendyala, R., *Forecasting Statewide Freight Toolkit*, Publication NCHRP Report 606, Transportation Research Board, 2008.
(2) Cambridge Systematics, Inc., *Validation and Sensitivity Considerations for Statewide Models*, Publication NCHRP Project 836-B Task 91, American Association of State Highway and Transportation Officials Standing Committee on Planning, September 2010.
(3) *Quick Response Freight Manual II*, Publication FHWA-HOP-08-010, US Department of Transportation Federal Highway Administration Office of Freight Management and Operations, September 2007.
(4) Schiffer, R.G., Shen, H., Wu, Y., Kaltenback, K.D., and Rossi, T.F., A Tiered Approach to Validating the Integrated Florida Statewide Model, Presented at the 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
(5) Gunn, H., Burge, P., and Miller, S., The validation of the UK National Transport Model: a backcasting approach, Proceedings of the European Transport Conference, Strasbourg, France, September 2006.

(6) Cambridge Systematics, Inc., *FSUTMS-Cube Framework Phase II: Model Calibration and Validation*, Prepared for Florida Department of Transportation Systems Planning Office, October 2008.