Development of a Freight Demand Model with an Application to California

Robert Lim[1,*], Zhen (Sean) Qian[2] and H.M. Zhang[3]
[1]San Francisco Municipal Transportation Agency, San Francisco, CA, 94103
[2]Heinz College, Carnegie Mellon University, Pittsburgh, PA, 15213, Tel: 412-268-7202
[3]Department of Civil and Environmental Engineering, University of California, Davis, Davis, CA, 95616

ABSTRACT
This paper discusses the disaggregation of the Federal Highway Administration's Freight Analysis Framework (FAF) database (version 3.0) on freight origin-destination data and the development of linear regression equations to describe the relationships between commodity-based freight trip productions/attractions to specific economic variables. Instead of generating a production/attraction equation for each commodity, commodities are grouped in certain ways to simplify model development and application. We consider three grouping methods and two model selection criteria (with and without intercepts), which are compared in terms of goodness of fit with two data sets (FAF versions 2.0 and 3.0). Furthermore, the freight generation models are validated using county-level economic data in California and applied to predict year 2015 commodity outputs. The results of this study can help city, county, metropolitan and state level planning agencies develop their own customized freight demand generation models without performing costly large-scale surveys.

Key Words: Freight transportation; Trip generation model, FAF3; Commodity flow

1. INTRODUCTION
Demand in freight transportation and the movement of goods continues to rise with the increase in population at state, domestic and global levels. Increased freight demand brings challenges such as added stress on already congested transportation networks

[*]Corresponding Author Email Addresses: Robert.Lim2@sfmta.com (Robert Lim), seanqian@cmu.edu (Zhen (Sean) Qian), hmzhang@ucdavis.edu (H.M. Zhang)

and negative impacts to air quality. In order to address these challenges there needs to be a greater understanding of freight and its impacts on the transportation network.

Many sources exist to describe goods movement but information available at the public level is mostly aggregated to the state level. The Federal Highway Administration (FHWA) maintains the Freight Analysis Framework database (FAF3, released in 2010) to estimate commodity flows and related freight transportation activity among states, regions, and international gateways. While useful, the data needs to be disaggregated to a finer level in order to perform an analysis on state and county-level goods movement. Once disaggregated, the FAF3 database in combination with other data sources will be used to develop a model to predict the demand for specific commodities by the county or state level, similar to that of a trip generation model.

Since disaggregation of national-level data to county-level data is necessary to produce meaningful and accurate predictions at these lower levels, much research has been done on the disaggregation process [11, 14]. In addition, there exists a private firm that specializes in this research, IHS Global Insight. This company maintains the TRANSEARCH database, which describes freight flows across the US at the national, state, and county levels. However, the use of such service is costly. Also in a comparison of the TRANSEARCH database with other disaggregation methods, the data do not match up [11].

This paper establishes freight demand generation models for trucks using the FAF3 database and other publicly available data sources. The proposed models will allow users to predict how much of a specific commodity in tonnage is produced (exported) from or attracted (imported) to any region (for example, city or county). This model by no means provides a "one-size-fits-all" freight generation model. Instead, this model provides insights on how one can go about creating their own version with data validation and how to adapt the model to the region for which the analysis is being done. Three models depending on the grouping of commodities that are carried by trucks will be presented in this paper, but there will not necessarily be an argument for which model is "better". The models differ from one another by how the commodities are grouped together, in which way the freight demand may be validated and an analysis of how the results are affected by these different groupings is provided.

A better understanding of goods movement across the network can be achieved via data analysis. By knowing what the demand of goods is to the specific location along with trucking characteristics such as load factors per good, truck flows can be mapped across the network. The information of where trucks go, and how often they go and what they are transporting can allow us to improve goods movement efficiency. The problem however lies with knowing what type of data to acquire and from what sources.

The idea for this paper is refined through a review of current practices in the goods movement field. Freight-specific generation models are often developed using methodologies that are applied to passenger forecasts [13]. In general, there are two approaches to truck trip generation estimation, commodity-based and trip-based [6]. The commodity-based method involves estimating the commodity flow tonnage and then converting it to the number of truck trips using a payload conversion factor.

Washington State University developed truck payload conversion factors to go from tonnage to trips through a series of roadside truck surveys [8]. Previous research has used economic variables such as employment, employment density, and floor space to assess the amount of commodity entering or leaving a given area. Commodity flows between given locations are then allocated to traffic analysis zones.

The second approach to estimate truck trip generation rates is through the trip-based method. Trip generation is developed through land-use data, roadside counts and surveys. The trip-based approach is typically evaluated on a site-by-site basis while the commodity-based approach is applied at the regional and zonal level. For this paper the commodity-based approach will be used because the regional and zonal-level economic data is easier to obtain, and easier to adapt to our needs compared to the site-specific information of the trip-based approach.

The extent of freight generation research ranges from predicting commercial delivery rates in metropolitan areas such as Melbourne [10], New York [2] and Chicago [15] to seaport container movement [9]. Our research will predict rates for movement of goods in and out of counties in California. In addition, future freight demand will be predicted. Historically, the simplest and most direct method to forecast future freight demand is to use existing freight demand data [3]. However, in order to make any prediction rates at all, the FAF database needs to be disaggregated.

As mentioned earlier, economic data and forecasts of industrial output and consumer demand are used to estimate annual production and consumption of goods. This data is provided at the national level through the Census Bureau, and then disaggregated to the state level. State level data is further disaggregated using local employment data to counties, cities, and traffic analysis zones. Data disaggregation of national-level data to county-level data is necessary to produce meaningful and accurate predictions at these lower levels. Much research has been done on the disaggregation process but there also exists a private firm that specializes in this research, IHS Global Insight. This company maintains TRANSEARCH database which describes freight flows across the US at the national, state, and county levels. However, the use of such service is costly. Also in a comparison of the TRANSEARCH database with other disaggregation methods, the data do not match up [13].

Two states, New Jersey [11] and Florida [14], along with the private firm Cambridge Systematics [4, 5] have taken similar approaches to disaggregating the commodity-based data of the FAF database, albeit it being of the older 2.0 version. This paper will disaggregate the current database version 3.0 (FAF3). The most frequently used approach uses some form of employment and population as the primary factors for commodity flow disaggregation. Mathematical relationships are developed between these factors and the actual tonnage origin and destination values of each commodity, as classified by the Standard Classification of Transported Goods (SCTG) system. Whereas regression equations have been typically developed per commodity (42 classified types in total), Florida's disaggregation approach includes combining similar commodities together into groups. Regression equations are then developed per each grouping.

This paper contributes to the literature of freight demand generation model in the following aspects. We propose a new commodity grouping method following the works

of [4, 5, 14], with however, new selection criteria. The up-to-date FAF3 database is used to build these models. In addition, we also apply Florida's and Cambridge Systematic's groupings to the same data sources, and compare all three models in terms of the generated outputs of the production and attraction tonnage commodity values. The extent of validating these models in previous studies, compares the R2 values of the generated linear regression equations with that of the TRANSEARCH values. In this paper, FAF data is used in both cases but the difference is that the model is calibrated with national-level data and validated by estimating county-level trip generation and then aggregated up to compare against state-level FAF numbers.

The remainder of this paper is organized as follows. Section 2 focuses on how the data is acquired and the process undertaken to transform the data into a usable form. Section 3 goes into the model development method. The results obtained from the regression models are discussed in Section 4, while Sections 5 and 6 detail how the model output is validated and the comparison of the model output to the existing FAF3 values for the base year and future year. Section 7 concludes the paper.

2. DATA ACQUISITION

The production of a model to describe goods movement requires the gathering of data from a variety of sources. Table 1 displays the type of data acquired for model development.

The initial process to develop a model to describe freight flows requires the disaggregation of the current Freight Analysis Framework (FAF) database. This database is sourced primarily from the 2007 Commodity Flow Survey (CFS). The FAF3 dataset is sectioned into 123 domestic geographical regions, five of which are in California: Los Angeles, Sacramento, San Diego, San Francisco, and the remainder of California. The previous version of the FAF database (FAF2, released in 2006) contained 114 zones.

FAF3 contains information on the tonnage of goods moved from one region to another for each of the commodity types for years 2007, 2015, 2020, 2025, 2030, 2035, 2040. FHWA considers 2007 as the baseline year and makes predictions for future years using a variety of growth factors. Commodities being moved out the origin FAF3 region are classified as a "production" value while commodities with a destination of a specific FAF3 region are classified as an "attraction" value. Instead of normally associating productions and attractions with the number of vehicle trips, this model will measure productions and attractions in terms of tonnage for a specific commodity. The 2007 data is used to build a tonnage production and attraction model for each type of commodity.

The preferred method to generate production and attraction values in goods movement is through the linear regression models. Productions and attractions are the dependent variables. The independent variables are a combination of population, employment, farmland, crop sales, and energy data. Table 1 displays the type of data acquired for model development. Since the FAF3 database uses 2007 data, explanatory variable data sources are acquired at years that are as close as possible to this year.

Table 1. Data used for linear regression model development

Data Type	Units	Source	Base Year	Future Year
Commodity Productions	Tons	FHWA Freight Analysis Framework 3 (FAF3)	2007	2015
Commodity Attractions	Tons	FHWA Freight Analysis Framework 3 (FAF3)	2007	2015
Population	Persons	U.S. Census Bureau	2008	
		California Department of Finance2008		2015
Employment	Employees	U.S. Census Bureau County Business Patterns	2008	
		California Employment Development Department		2006-2016 or 2008-2018
Farmland	Acres (1000s)	USDA: The Census of Agriculture	2007	
		Caltrans: California 2008-2030 County-Level Economic Forecast		2015
Crop, Livestock Sales	$US (1,000,000s)	USDA: The Census of Agriculture	2007	
		Caltrans: California 2008-2030 County-Level Economic Forecast		2015
Net Annual Electrical Generation (Coal)	Megawatt-hours (Mw-h)	National Energy Technology Laboratory (NETL) Coal Power Plant Database (CPPDB)	2005	
		U.S. Energy Information Administration Coal Supply, Disposition, Prices, Projections to 2035		2015

The Standard Classification of Transported Goods (SCTG) system [1] is used to categorize specific types of goods (also referred to as commodities). Commodities are allocated into 43 classifications as numbered in Table 2.

Population data come from the U.S. Census Bureau. This data contains the population count in the United States as of July 1st for the years 2000 to 2008. For consistency, the latest available dataset, July 1, 2008 is selected to match the 2008 County Business Patterns (CBP) database. The CBP database is used for employment data per specific job classification and follows the North American Industry Classification System (NAICS).

Table 2. Commodity grouping set II

Category	Commodity (SCTG #)	Category	Commodity (SCTG #)
A (food)	Animals and Fish (1)	F (lumber)	Logs (25)
	Animal feed (4)		Wood prods. (26)
	Meat/seafood (5)		Newsprint/paper (27)
B (agricultural products)	Cereal grains (2)	G (paper)	Paper articles (28)
	Other ag prods. (3)		Printed prods. (29)
C (consumables)	Milled grain prods. (6)		Textiles/leather (30)
	Other foodstuffs (7)	H (manufactured goods)	Furniture (39)
	Alcoholic beverages (8)		Misc. mfg. prods. (40)
	Tobacco prods. (9)		Mixed freight (43)
D (materials)	Building stone (10)		Base Metals (32)
	Natural sands (11)	I (metals)	Articles-base metal (33)
	Gravel (12)		Machinery (34)
	Nonmetallic minerals (13)	J (electronics)	Electronics (35)
	Metallic ores (14)		Precision instruments (38)
	Basic chemicals (20)	K (motorized vehicles)	Motorized vehicles (36)
	Pharmaceuticals (21)		Transport equip. (37)
	Fertilizers (22)	L (waste)	Waste/scrap (41)
	Chemical prods. (23)		
	Plastics/rubber (24)		
	Nonmetal min. prods. (31)		
E (fuel)	Coal (15)		
	Crude petroleum (16)		
	Gasoline (17)		
	Fuel oils (18)		
	Coal-n.e.c. (19)		

Production and attraction equations fit poorly against farm employment numbers [4]. A different method using farm acreage is applied instead to generate production and attraction equations. Farm acreage data, listed by state and county, is acquired from the United States Department of Agriculture's (USDA) Census of Agriculture. Farm acreage is listed by state and county.

The two data categories of crops and livestock sales, in units of U.S. dollars, are also acquired from the same USDA source. Crop and livestock sales can show how productive farms are, which can have a better correlation with employment numbers than farm acreage.

Equations for coal (SCTG 15) are developed with data from the U.S. Department of Energy's National Energy Technology Laboratory (NETL) as an explanatory variable.

The NETL's Coal Power Plant Database (CPPDB) lists the net annual electrical generation of each domestic coal power plant in megawatt-hours (MWh). Since information on the plant location's county is available, a FIPS state and county code is assigned to the data value.

In addition to the state-level data, we also collected county level data for CA, as this will develop a model applied in all counties in CA. The year 2015 is chosen as the comparison year for projection data because it is the first available set of projected data from FHWA. The 2015 FAF3 data is primarily based on the growth (or decline) in the demand for the commodities while the 2015 explanatory data is essentially based on the growth (or decline) in employment categories that produce those commodities. Table 1 also lists the year of the resources used to project our data for the year 2015.

The U.S. Census Bureau does not make predictions for population growth at the county level. However, the California Department of Finance does produce county level population for 2015. Because we originally used Census data to develop our model and the California dataset is slightly different numbers-wise from the census data, we developed a method to apply the projections in population predicted by the California Department of Finance to the Census data. The percentage difference in growth in a particular county from the year 2008 to the year 2015 is applied to the 2008 Census population value for that county to produce the population figure for 2015 that we will implement into the model equations. The population equation is applied to each county in California (CA).

$$Projected\ Population_{2015} = Pop_{2008\ Census} + \left(Pop_{2008\ Census} * \frac{Pop_{2015\,CA} - Pop_{2008\,CA}}{Pop_{2008\,CA}} \right) \quad (1)$$

Since the Census Bureau also does not make employment projections per NAICS category at the county level, projected growth/decline figures from the California Employment Development Department is applied to the 2008 Census CBP data. This method is applied in the similar fashion to that of our population projection, where the percentage difference for employment values between the baseline year and the chosen projected year of the CA EDD data are taken and applied to the 2008 Census CBP dataset. However, the CA EDD dataset comes in two forms, (1) projections to the year 2016 from 2006 baseline data and (2) projections to the year 2018 from 2008 baseline data. To acquire 2015 values from these datasets, we assume that absolute changes in employment follow a linear pattern. Therefore for the 2006-2016 dataset, we take 9/10 of the actual value for the selected employment category since there are 9 years between 2006 and 2015. Likewise with the 2008-2018 dataset, we take 7/10 of the actual value for that employment category since there are 7 years between 2008 and 2015. From then on, we can apply the equation below to generate a value for employment for a particular NAICS category in a certain county.

$$Projected\ Employment_{2015} = Emp_{2008\ Census} + \left(Emp_{2008\ Census} * \frac{Emp_{2015\,CA} - Emp_{2008\,CA}}{Emp_{2008\,CA}} \right) \quad (2)$$

Employments values are projected for each NAICS category that we use in our regression equations for each of the 58 California counties. Depending on the size of the county (population and employment wise), projections for all employment categories to the 3-digit NAICS detail level are made only for the "larger" counties such as Los Angeles and San Francisco. When there is no projection to the 3-digit level for a particular NAICS value, the percentage difference from 2008 to 2015 of the 2-digit level NAICS category is used. For example, if there is no value for a particular county in the 3-digit NAICS category of "Merchant Wholesalers, Durable Goods" (423), the value for the 2-digit NAICS category of (42) "Wholesale Trade" will be used instead.

For one specific NAICS category, "Support Activities for Agriculture and Forestry" (115) there was neither a projection made for each individual county and for California as a whole. As a result, we have to assume that there is zero growth from 2008 to 2015 in that category.

Since there are no datasets with future projections for farmland acres and crop/livestock sales, other economic variables are used to make such projections; one variable from this forecast, crop value, is used for our projection purposes for the farmland, crop sales and livestock sales variables. We make the assumption that growths or declines in crop/livestock sales are correlated with the corresponding growths and declines in a county's available farmland. In certain regions, the surge in population may require local governances to take away farmland by building housing and retail space. On the other hand, certain regions may just do the opposite by expanding farmland as a means of curbing growth. Since this land use issue is beyond our scope of study, we will follow the assumption made earlier.

This Caltrans forecast projects the crop value for each county for calendar years 2008 to 2030. Since these three agricultural variables are taken from a different source in the model development stage, the USDA, we will use the percentage in growth or decline in crop value from 2007 to 2015 and apply the value to the 2007 USDA value for each county. The year 2007 is chosen, instead of 2008 like the population and employment variables because the USDA data was for the year 2007. The percentage change in crop value will be applied across the board to farmland, crop sales and livestock sales in each county and will follow the formula below to generate a value for each of the three variables. In this case, 'X' denotes the variable being projected (farmland or crop sales or livestock sales).

Projected Farmland or Crop Sales or Livestock Sales$_{2015}$ =

$$X_{2007\,USDA} + \left(X_{2007\,USDA} * \frac{Crop\,Value_{2015\,Caltrans} - Crop\,Value_{2007\,Caltrans}}{Crop\,Value_{2007\,Caltrans}} \right) \quad (3)$$

Data from the U.S. Energy Information Administration (U.S. EIA) is used to project coal generation data in 2015. Projections for coal production are made for every year from 2008 to 2035. In this case, the percentage change in the total domestic 2015 coal production value from the total domestic 2008 coal production value is calculated. Since our original coal data is based on 2005 numbers, we apply a factor of 10/7 to the percentage change value as there is a 10-year difference between the 2005 dataset and

the 2015 projection year and a 7-year difference between the 2008 and 2015 projection years of the U.S. EIA dataset. This new value, -15%, is applied uniformly to those California counties that are reported to have coal generation in the NETL database.

3. MODELS

The typical method to develop a commodity production and attraction generation model is to associate employment by industry and the commodities those industries produce and consume. While regression equations can be developed for each commodity type, it is not efficient to produce many equations when certain commodities are similar and can be grouped together. The grouping of commodities is necessary to obtain a statistically significant industry category that is realistically involved in the production or attraction of the specific commodity. Commodities can be grouped together based on factors such as that certain geographic regions produce and attract specific commodities, employees of homogenous skill levels work in these geographic regions, or that the specific commodities are often transported together. For example, the commodities motorized vehicles (SCTG 36) and transportation equipment (SCTG 37) can be aggregated together because they are typically produced in similar geographic regions and are attracted to similar regions.

Three sets of groupings are presented. Grouping I follows the arrangement used in Cambridge Systematics' report [5]. Grouping II consists of our own arrangement of the commodities. Grouping III adheres to the state of Florida's FAF disaggregation model [14].

3.1. Commodity Grouping I (Model I)

Cambridge Systematics' method for developing regression equations for productions is different from their method for attractions [4]. In the case of attractions, separate regression equations are developed for each of the 42 commodities. As for productions, 30 regression equations are generated for 30 separate commodities while 3 equations are generated for the 3 aggregated groups of commodities. Most of the production of commodities in a geographic region is a function of the associated industry employment for that commodity. For example, the production of paper goods should be associated with the number of employees in the paper industry in that region. However, the attraction of paper goods is not necessarily related to paper industry employment levels.

Cambridge Systematics' report was published in 2009 using FAF2 data. With the release of the FAF3 database, Cambridge Systematics' methodology will be applied in this paper to the up-to-date database.

As the FAF2 database uses 2002 data as a baseline and the FAF3 database uses 2007 data as a baseline, the comparison between these datasets will tell us what has changed in this five year span in terms of the relationship between employment industries and the commodities that they produce and consume.

3.2. Commodity Grouping II (Model II)

A downside to aggregating commodities together is that the groupings may not be truly reflective of the geographic regions. For example, if one groups two commodities together, one region may be a large producer of one of the commodities but not of the

other. Therefore when it comes to validating or applying the model, model generated tonnage values for certain commodities can be drastically different depending on the commodity aggregation. Consequently it is very important to place commodities into "correct" groupings. But since there is no one-size-fits-all method to group these commodities, two groupings are presented in this paper (aside from Cambridge Systematics' groupings). The first commodity grouping set is model II, listed in Table 2.

These 12 groupings are based upon the upon the U.S. Department of Commerce's description of each commodity type [1]. Commodity groupings are based on the following: (1) they have skill-wise similar workforces to produce the good (2) are produced in similar geographic regions and (3) are attracted to similar retail selling venues. Due to these reasons, it is good practice to group commodities differently and separately for productions and attractions. For example, California may not produce much of the gas and fuel related commodities of group E in Table 2, but they are large consumers (attractors). Of course, the best grouping is based on knowledge of local production and consumption characteristics.

3.3. Commodity Grouping III (Model III)

Florida's commodity disaggregation of the FAF database utilizes 13 groupings [14]. The most distinguishable differences between groupings II and III are that non-durable goods and durable goods are aggregated into only two distinct groups in the Florida model. In grouping II, commodities that are identified as a non-durable or durable goods are not necessarily restricted to being categorized into these two groupings. For example, goods of category J (other durable manufacturing) of model III are placed among categorical groupings of materials, metals, electronics, manufactured goods, and motorized vehicles. The general definition is that a durable good lasts for a long period of time and can be considered as a form of investment spending [12]. Thus a non-durable good has a shorter life expectancy. But due to many differences (use of the good, method of production, etc.) between goods within either category, it is more applicable to group commodities that follow our model II convention.

3.4. Model Development

Linear equations were fitted to the annual tonnage for each commodity or commodity group that was provided by the FAF3 database to the selected explanatory variables. To generate a tonnage production or attraction value for a grouped commodity, the individual production or attraction value for each commodity within the group is added together. The explanatory variables include economic and geographic indicators such as population, employment by category, farm acreage and crop sales, as discussed in the section of data acquisition.

All 123 FAF3 zones are chosen as data points in the regression as opposed to using the five California FAF3 zones as data points. This method is chosen to allow for greater confidence in the relationships between production/attraction tonnage values with the explanatory variables. The assumption will be that the developed commodity regression equations with national level data will be applicable to California for the validation purposes. The downside to this assumption is that some industries that are

not as prevalent in California compared to other states (e.g. coal) will be applied to the California model despite the application of national data based regression equations. A difficult trade-off is faced in that if only the five California FAF3 zones were used as data points, the regression fitting would be more California specific. However, this decision would come at the expense of greater relationship confidences when using the 123 data points of the national dataset.

Previous studies (i.e. Model I) on FAF database disaggregation calculate linear regressions using zero intercepts based on the assumption that a region with zero employment in an industry would not produce or attract any freight for the associated commodity. All three models follow this assumption but Model's II and III will also attempt to produce linear regressions with a constant in the equations to see if this assumption holds any true validity.

The number of explanatory variables used is dependent on each variable's statistical significance. Cambridge Systematics' approach was to keep explanatory variables if they were statistically significant at the 95 percent confidence level in most instances. However if a variable is judged to be critical in explaining the commodity production or attraction, it would still be included in the regression equation despite not being significant at the 95 percent level. When we apply Model I to FAF3 database in this research, we still adopt the criteria of keeping variables that are statistically significant in most cases since the model is essentially a upgrade of the model produced by Cambridge. On the other hand, models II and III keep explanatory variables in the regression equation only if they are significant at the 95 percent confidence level. The reasoning behind this is due to the fact that models II and III use commodity groupings, only variables that are significant should be kept because of the presence of multiple commodities. A variable that is not statistically significant may not accurately describe the relationship between the production/attraction of other commodities in that grouping.

4. RESULTS AND DISCUSSIONS

Here we discuss the results obtained from the regression models of all the three models.

4.1. Model I: FAF2 vs. FAF3

We first compare the results from Model I with FAF2 and FAF3 data. The focus of this comparison between the two years data is to see how the relationships of explanatory variables and tonnage values change over time. The output of this model can only be as good as the inputs into the model. If methods to calculate production and attraction tonnage values change between the two data years for any commodity, it is difficult to make a direct comparison. However, if these calculation methods do not change, we can determine if the explanatory variables used in 2002 can still accurately predict tonnage outputs in 2007 through the analysis of the coefficient and correlation values.

For a majority of the commodities, explanatory variable coefficients and R^2 values are very similar between the two sets of data years. However, certain commodities such as those in the coal and fuel sectors have noticeable differences in the variable coefficient values and the R^2 values between the 2007 and 2002 data as shown in Table 3. According to the FAF3

website, version 3.1 of the database was revised to incorporate improved estimates for imports of crude petroleum and refinement [7]. In some cases, the production and attraction values more than doubled for certain FAF3 regions for these specific commodities.

Another example of a significant difference is SCTG 35, electronics, shown in Table 3. The explanatory variables for the 2002 data produce a R^2 value of 0.70 while the 2007 data has a R^2 value of 0.34. The relationship between these variables (manufacturing categories) and the actual tonnage values is weaker than what was initially believed to be a strong relationship. Explanatory variables used in 2002 may not necessarily provide similar coefficients for the 2007 data. Further refinement of this model would remove explanatory variables that are not significant at the 95 percent confidence level. Since Model I was an exact replication of the Cambridge model, variables that are not significant at the 95 percent confidence level are still kept in the regression equation.

Different production and attraction tonnage calculation methods can have a major effect on the coefficient and correlation values. Since the method to produce a linear regression is the same, and the same explanatory variables are used in each regression equation, changes in the input data result in the difference of coefficient values for certain commodities between the two data years. However for the rest of the SCTG categories, there is not a significant change between years 2002 and 2007.

Table 3. Production regression equations for coal and fuel related commodities and electronics

Variable Description	FAF^2		FAF^3	
	Coefficient	T-Stat	Coefficient	T-Stat
SCTG 16 - Crude Petroleum				
Oil and Gas Extraction	8.324	5.36	0.031	11.23
R^2	0.21		0.51	
SCTG 17 - Gasoline				
Petroleum and Coal Products Manufacturing	7.592	23.67	3.543	10.56
R^2	0.83		0.48	
SCTG 18 - Fuel Oils				
Petroleum and Coal Products Manufacturing	3.885	19.39	1.749	7.32
R^2	0.77		0.31	
SCTG 35 - Electronics				
Machinery Manufacturing	0.02	3	0.013	1.13
Computer and Electronic Product Manufacturing	0.012	4.35	0.02	2.85
Electrical Equipment, Appliance, and Component Manufacturing	0.029	2.44	0.036	1.09
R^2	0.7		0.34	

4.2. Model II and Model III

Results of the R^2's using Model II and Model III are presented in Table 4. Since direct comparisons of the R^2's of model II and III groupings cannot be made, the tables are not presented side by side.

For both models, the R^2 values for regressions without a constant (zero intercept) are higher for every commodity in each production and attraction regression equation. The attraction regression R^2 values are also typically higher than that of the production R^2 values. In general, model III's R^2 values are lower than that of model II. This indicates that the groupings of commodities in Model II may explain more variations with respect to those explanatory variables, and may be more reliable in general. However, this does not necessarily mean that Model II is always "better" and the groupings of Model III are actually more reflective of how commodities are generally shipped together.

Regression equations that contain a constant produce a lower R^2 value because not all FAF zones produce or attract all of the commodities in their respective group. Therefore by forcing the regression equation to have a constant, a tonnage value will be assigned to that FAF region even if corresponding explanatory variables say otherwise. For instance, the production value for the food commodity grouping can be explained by the existence of food manufacturing employees in the FAF zones. If a FAF zone has no employees in the food manufacturing sector, a regression equation with the constant would provide a food production tonnage value. It is sometimes reasonable to have a production or attraction that is not directly related to the selected explanatory variables, thus a regression equation without a constant can provide a statistically more reliable

Table 4. R^2 comparisons of Models II and III

	Model II				Model III			
	Productions		Attractions		Productions		Attractions	
SCTG	With Constant	Without Constant	With Constant	Without Constant	With Constant	Without Constant	With Constant	Without Constant
A	0.77	0.84	0.83	0.89	0.77	0.82	0.57	0.68
B	0.75	0.81	0.73	0.8	0.05	0.23	0.6	0.83
C	0.63	0.8	0.92	0.96	0.38	0.41	0.55	0.57
D	0.58	0.81	0.71	0.87	0.73	0.85	0.92	0.96
E	0.46	0.64	0.75	0.86	0.22	0.4	0.8	0.88
F	0.74	0.81	0.77	0.84	0.74	0.81	0.77	0.84
G	0.58	0.75	0.86	0.92	0.29	0.53	0.64	0.82
H	0.55	0.73	0.89	0.95	0.52	0.69	0.84	0.91
I	0.26	0.53	0.78	0.88	0.35	0.54	0.76	0.88
J	0.21	0.38	0.79	0.88	0.29	0.52	0.78	0.89
K	0.34	0.49	0.68	0.77	0.42	0.69	0.65	0.81
L	0.59	0.76	0.75	0.85	0.59	0.76	0.75	0.85
M	–	–	–	–	0.53	0.71	0.8	0.9

description of a FAF zone, unless the zero intercept is necessary for certain commodities (petroleum products for example).

4.3. Grouping Choices

Certain commodity groupings will have higher R^2 values than others by nature, for example, the explanatory variable of population can more accurately describe the production/attraction of food commodities compared to coal. Not all U.S. regions produce or attract coal, but all regions will attract food, if not also produce. Another reason to explain the differences among commodity R^2 values are shipping logistics. Regions with seaports that import international goods, of which are then transported to trucks, classify the goods as productions since they originate from the region to which the seaport belongs. Therefore production values for commodities such as electronics will be skewed towards regions with seaports such as Los Angeles or Oakland. Since we are only analyzing truck movements, goods that are primarily moved by other modes (rail, air, sea), will be missed by this model. The total amount of goods produced and attracted in a region includes goods shipped by all modes. Therefore if a certain type of good is mainly transported through trucks, the model will work well. For other goods, such as coal, the models will not work well and the regression equations will be off.

In general, R^2 values are dependent on the grouping, the choice of explanatory variables, and the commodities involved. Due to this relationship, it is difficult to produce universally valid models with the available data.

5. VALIDATION

In our validation, the model is calibrated with FAF data and validated by estimating county-level trip generation and then aggregated up to compare against state-level FAF numbers. Each developed regression equation is applied to the 58 counties in California and then aggregated back up to the five California FAF3 zone levels. The reason for applying the equations at the county level instead of at the FAF3 zone level is to see the quantity of commodities produced and attracted at this finer data resolution. If an end user of these models wanted to acquire county level data, they would be able to do so with the developed regression equations. However, since we do not have county level data to validate against, the regression-based data has to be aggregated back up to the FAF3 zone levels.

The same data used to produce the regression equations are also used for validation purposes. This analysis will be made at the state level, since comparing the percentage difference of each commodity or commodity grouping for each of the five California FAF3 zones would be unnecessary for this step as we are looking at how well our model did overall.

5.1. Model I

Aggregating the validation results of the five California FAF3 zones produces a value for the entire state to compare against the actual database values (see Table 5(a)). The SCTG commodities with the most accurate values for productions and attractions are commodities 8 (alcoholic beverages), 29 (printed products), and 43 (mixed freight).

These commodities have production tonnage value differences ranging from -5.14% to +7.65% of the actual database values. As for attraction tonnage value differences, these commodities were within -5.11% to +11.61% of the actual database values.

Other commodities have percentage differences of up to +1450% (SCTG14 – metallic ores) of the actual value. The varying percentage differences is a result of developing regression equations based on the national data of 123 FAF3 data points. Because of the substantial geographic and economic differences among the 50 states, let alone counties, this one set of developed regression equations cannot accurately provide production and tonnage values that are close to the actual values for every FAF3 zone. As the other models will attest, there is no "one-size-fits-all" model to estimate production and attraction values.

5.2. Model II

The validation results of the regression equation based production and attraction data for Model II is presented in Table 5(b). The regression equations without a constant fared much better when matched against the actual FAF3 data. Validation results for equations without a constant range from being within -43% of the actual data to +93% of the actual data. Unlike some of model I's results where the percentage differences for certain commodities were off by more than 100%, the results from model II do not stray away as much from the actual data. This finding can be a result of the increased volume in tonnage produced and attracted when analyzing a grouping of commodities as opposed to a single commodity. Certain commodities within the groupings can over/undercompensate volume-wise for other commodities within the category. For instance in grouping D, the volume of gravel (SCTG 12) can be much greater than the volume of all the other commodities put together. As a result, the validation results can be skewed to where gravel has a larger influence on the validation results compared to the other commodities within that grouping.

5.3. Model III

The validation results of the regression equation based production and attraction data for Model III is presented in Table 5(b). Similar to the validation results of model II, equations that do not contain a constant value generated results that were more in line with the actual FAF3 data. However, the one commodity grouping validation result that stands out is that of grouping C, which is composed of a single commodity, coal. Since these regression equations are only validated with California FAF3 data, certain commodities and industries that are present nationwide do not hold as significant of an influence in this state comparatively. California is not a coal producing or attracting state, yet the census data show that there are employees in the categories used to predict the tonnage values. The regression equation for the production of coal is based on employment in the NAICS categories of 211 (oil and gas extraction) and 212 (mining (except oil and gas)). California does indeed have employees in these two categories, yet they do not factor into the production of coal as much as in the other FAF3 zones. Coal is not necessarily shipped by truck, but by rail primarily. Because our freight models deal only with truck movements, we are not able to describe coal movement in its entirety and

Table 5. Validation results for the year of 2007

(a) Model I, % Difference From Actual 2007 CA FAF[3] Values

SCTG	Regr. Eq., w/o Constant		SCTG	Regr. Eq., w/o Constant	
	Productions	Attractions		Productions	Attractions
1	207.16%	62.34%	22	10.85%	3.03%
2	414.80%	429.69%	23	10.85%	3.31%
3	-13.07%	9.14%	24	-14.01%	3.88%
4	84.47%	112.52%	25	-60.14%	46.68%
5	13.46%	3.22%	26	-22.36%	4.99%
6	-21.68%	-8.26%	27	29.91%	-6.26%
7	-24.26%	1.82%	28	-13.20%	1.47%
8	7.65%	-5.11%	29	3.21%	11.61%
9	182.90%	139.18%	30	-66.07%	-13.80%
10	-73.31%	44.10%	31	-23.55%	185.95%
11	-73.31%	2.26%	32	25.04%	623.72%
12	-73.31%	101.20%	33	5.48%	19.48%
13	-73.31%	149.27%	34	-22.65%	10.10%
14	-73.31%	1450.64%	35	-9.12%	-14.77%
15	-73.31%	956.42%	36	-27.62%	-9.86%
16	-81.38%	-18.00%	37	-27.62%	71.93%
17	-40.18%	-24.34%	38	-30.68%	-36.75%
18	74.29%	467.52%	39	-34.90%	-11.13%
19	-28.03%	13.31%	40	20.54%	10.84%
20	10.85%	697.46%	41	-93.97%	-19.16%
21	10.85%	-9.18%	43	-5.14%	4.42%

(b) Models II and III % Difference from Actual 2007 CA FAF[3] Values

SCTG	Model II				Model III			
	Productions		Attractions		Productions		Attractions	
	With Constant	Without Constant	With Constant	Without Constant	With Constant	Without Constant	With Constant	Without Constant
A	103.96%	48.82%	75.13%	17.53%	194.37%	90.57%	-6.67%	36.19%
B	186.78%	89.83%	142.62%	93.21%	649.72%	-83.89%	209.31%	42.11%
C	55.10%	-5.75%	-22.16%	-2.82%	-29943.13%	6704.16%	-38834.73%	3058.14%
D	172.22%	14.73%	179.62%	46.91%	64.39%	-5.00%	-1.93%	0.90%
E	101.17%	-43.14%	-2.41%	-0.98%	48.11%	-13.11%	-10.27%	-27.80%
F	138.00%	-9.83%	57.50%	-3.48%	141.53%	-9.82%	57.75%	-3.49%
G	80.21%	-8.16%	8.17%	4.48%	306.36%	10.83%	103.01%	20.13%
H	76.72%	15.91%	23.48%	5.89%	100.29%	-11.70%	6.17%	1.47%
I	269.92%	45.24%	86.45%	32.29%	192.64%	-31.88%	41.46%	9.59%
J	98.52%	-7.00%	-20.68%	-12.39%	157.58%	18.66%	36.31%	22.04%
K	65.98%	-27.99%	-97.09%	-6.71%	74.47%	-23.56%	72.49%	0.21%
L	-13.03%	-35.34%	-26.44%	-19.16%	-13.03%	-35.34%	-26.44%	-19.16%
M	–	–	–	–	54.55%	10.05%	10.76%	6.48%

the validation data reflects this. On the other hand, goods that are primarily moved by truck are likely to have better validation scores, but not necessarily because of other dependent factors such as geography and the specific type of good that is moved.

6. COMPARISON OF DEMAND FORECAST IN 2015

The 2015 model projection results are similar percentage difference wise to that of the 2015 FAF3 data for certain commodities and commodity groupings. Table 6(a) displays the 2015 data comparison for Model I.

Thirty-six of the 75 productions and attractions commodity data results improved upon their respective 2007 validation results. Improvement is measured by if the percentage difference between the 2015 model and 2015 FAF3 results are smaller than the percentage difference between the 2007 model and the actual 2007 FAF3 results. Even though 39 of the 75 data results were "worse" off, a majority of the percentage differences between the model and FAF results were within +/- 10 percent. This finding is promising because if the percentage differences between the commodity or commodity groupings of the two dataset years are similar, we can deduce that projections on future job demand in specific employment categories is related to the future demand in commodities.

The application results of models II and III are presented in Table 6(b). For model II, The percentage difference between the two dataset years is close for certain commodity groupings. Groupings that have high R^2 correlation values tend to have model results that are near the values of the actual FAF3 figures as explained in section 4. Also, those groupings with high correlation values are more likely to have similar percentage difference numbers between the two dataset years. The validation comparison results of Model III add further to the argument that a relationship exists between the demand in employment and the demand in commodities. Even though the results of grouping "C" of Model III show that our model results are completely inaccurate to those of the FAF3 dataset, it is interesting to note how very much alike the percentage difference between the 2007 and 2015 data are. Grouping "C" describes the commodity of coal, which California is neither a heavy producer nor attractor. Again, the percentage difference between the model and FAF3 data of other specific commodity groupings are very much similar.

Despite two distinct methods to project 2015 production and tonnage values for different commodities and commodity groupings, the results from our developed regression model show that there is a link in the growth (or decline) in employment in certain categories with the corresponding growth (or decline) in the commodities that these employees produce and attract. The employment variables in our model are based on provided projections from the California Employment Development Department for specific industries while the projections of the 2015 FAF3 data is based on future demand in the commodities themselves. The variables for each county are applied to our model, and then summed to the respective FAF3 region as we have done before. The percentage differences between the model results and the FAF3 data for specific commodities are compared for the two study years of 2007 and 2015. The groupings with high R^2 with their explanatory variables are more likely to have similar percentage

Table 6. Comparison results for the year of 2015

Model I % Difference From 2015 CA FAF[3] Values					
SCTG	Regr. Eq., w/o Constant		SCTG	Regr. Eq., w/o Constant	
	Productions	Attractions		Productions	Attractions
1	178.74%	36.05%	22	10.34%	21.29%
2	285.57%	296.63%	23	10.34%	-5.96%
3	-30.73%	-11.10%	24	-24.45%	-4.18%
4	33.45%	62.37%	25	-67.49%	26.63%
5	-6.14%	-9.44%	26	-20.77%	12.16%
6	-29.96%	-16.80%	27	9.21%	-12.76%
7	-36.20%	-9.41%	28	-29.94%	-7.59%
8	-2.24%	-12.06%	29	28.73%	44.89%
9	361.65%	337.76%	30	-73.89%	-18.80%
10	-70.25%	65.50%	31	-20.60%	204.73%
11	-70.25%	19.75%	32	10.26%	606.03%
12	-70.25%	121.17%	33	-3.64%	16.16%
13	-70.25%	158.01%	34	-33.44%	3.63%
14	-70.25%	1021.36%	35	-23.24%	-25.87%
15	-70.25%	886.52%	36	-37.38%	-14.31%
16	-84.63%	-34.07%	37	-37.38%	45.23%
17	-42.26%	-24.06%	38	-74.03%	-71.69%
18	58.85%	423.30%	39	-14.86%	0.62%
19	-37.79%	-3.18%	40	1.69%	0.65%
20	10.34%	50.31%	41	-94.32%	-16.34%
21	10.34%	-20.99%	43	-13.22%	-1.24%

(b) Model II and III % Difference from 2015 CA FAF[3] Values

	Model II				Model III			
	Productions		Attractions		Productions		Attractions	
SCTG	With Constant	Without Constant	With Constant	Without Constant	With Constant	Without Constant	With Constant	Without Constant
A	48.82%	17.53%	16.72%	-5.94%	90.57%	36.19%	51.19%	7.08%
B	89.83%	93.21%	50.78%	52.96%	-83.89%	42.11%	-81.48%	55.85%
C	-5.75%	-2.82%	-17.77%	-12.13%	6704.16%	3058.14%	6565.31%	3024.11%
D	14.73%	46.91%	21.59%	57.89%	-5.00%	0.90%	-20.47%	-12.62%
E	-43.14%	-0.98%	-47.29%	-1.30%	-13.11%	-27.80%	-16.78%	-25.02%
F	-9.83%	-3.48%	-11.58%	-5.45%	-9.82%	-3.49%	-11.57%	-5.46%
G	-8.16%	4.48%	-17.88%	3.79%	10.83%	20.13%	10.31%	20.15%
H	15.91%	5.89%	2.59%	1.74%	-11.70%	1.47%	-27.33%	-4.41%
I	45.24%	17.53%	31.88%	-5.94%	-31.88%	36.19%	-36.86%	7.08%
J	-7.00%	-12.39%	-35.56%	-30.26%	18.66%	22.04%	3.91%	15.05%
K	-27.99%	-6.71%	-37.71%	-10.29%	-23.56%	0.21%	-20.60%	10.28%
L	-35.34%	-19.16%	-39.66%	-16.34%	-35.34%	-19.16%	-39.66%	-16.34%
M	–	–	–	–	10.05%	6.48%	-4.25%	0.29%

difference numbers from the FAF data between the two dataset years. Additional future projections for commodities for different years through the use of our model can be made based on the availability of future projections in employment data.

7. CONCLUSIONS

We developed a freight demand model using disaggregated FHWA's FAF database and compared it with two other models. The first model builds upon the work of previous research by generating production and attraction equations for each commodity. The second and third models group similar commodities together in order to simplify model development and application. The grouping of commodities is sometimes necessary in order to obtain statistically significant relationships between the independent explanatory variables and the dependent production and attraction values.

Since each commodity has a production and an attraction value, a set of two linear regression equations are developed for each commodity. Data is sourced from two sets of years, 2007 and 2015. FHWA considers year 2007 to be the baseline year, as future projections for productions and attractions build upon the baseline data. Year 2015 data is also acquired because this specific year is the first in a set of projections that the FAF database provides.

The baseline year results of Model I show that there were no significant changes in correlation values for each SCTG commodity. For the most part, the R^2 values for the commodity groupings of Model II are higher than the R^2 values in other models. But this does not indicate that Model II is more applicable than other models and it can be argued that the groupings of Model III are actually more reflective of how commodities are generally shipped together. Certain commodity groupings will have higher R^2 values than others by nature, for example, the explanatory variable of population can describe better the production/attraction of food commodities compared to that of coal.

The production/attraction output of each equation is validated against the actual production/attraction value of the FAF3 database. The equations are applied to each of the 58 California counties to show the productions and attractions of each commodity in that specific county. Because the FAF3 data is at a regional level, the model generated county level data is aggregated to the regional level for comparison purposes. Certain model generated commodity output values are near the actual FAF3 value, while other values are off. Because of the substantial geographic and economic differences among the 50 states, let alone counties within California, the developed equations cannot possibly provide production and attraction tonnage values that are near the actual FAF3 values for every region.

The models were also applied to California for the year 2015. Despite two separate methods (FHWA's method vs. our method) to predict 2015 production/attraction values for different commodities and commodity groupings, the results from our developed models establish a link in the growth (or decline) in employment in certain categories with the corresponding growth (or decline) in the commodities that these "employers" produce and attract. The percentage differences between the model results and the FAF3 data for specific commodities are compared for the two study years and prove that commodities with equations of high R^2 values are more likely to have low percentage difference values between the two dataset years.

The generated data generated from these models can be very useful. Application possibilities include using the data to determine where infrastructure investment dollars should be allocated. Areas with high goods movement activity can be prioritized to receive more funds, which is a more equitable method compared to the current political based decision-making. Future work in this field may further analyze how commodities are transported between their origins and destinations, including the conversion of tonnage values to actual truck trips.

REFERENCES

[1] *2007 Commodity Flow Survey Standard Classification of Transportation Goods (SCTG): SCTG Commodity Codes*. U.S. Department of Commerce, Publication CFS-1200. 2006.

[2] Bastida, Carlos, and Holguin-Veras, Jose. 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. 167–175.

[3] Beagan, Daniel, Fischer, Michael J and Kuppam, Arun R. *Quick Response Freight Manual II*. Publication FHWA-HOP-08-010. FHWA, U.S. Department of Transportation, 2007.

[4] California Commodity Origin-Destination Database. *Technical Memorandum*. Cambridge Systematics. Jan 2010. Print.

[5] Cambridge Systematics. Development of a Computerized Method to Subdivide the FAF^2 Regional Commodity OD Data to County Level OD Data. Jan 2009.

[6] Fischer, Michael J., et al. Local Truck Trip Generation Data: A Synthesis of Highway Practice. In *NCHRP Synthesis*, No. 298, Transportation Research Board of the National Academies, Washington, D.C., 2001.

[7] Federal Highway Administration. *Freight Analysis Framework*. www.ops.fhwa.dot.gov/freight/freight_analysis/faf/index.htm. Accessed August 20, 2010.

[8] Gillis, William R., et al. "Movement of Freight on Washington's Highways: A Statewide Origin and Destination Study". *Eastern Washington Intermodal Transportation Study*, 1995.

[9] Holguin-Veras, Jose, Y. Lopez-Genao and A. Salam, Truck-Trip Generation at Container Terminals: Results from a Nationwide Survey. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1790, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 89–96.

[10] Ogden, K. W. Modeling Urban Freight Generation. Traffic Engineering and Control, Vol. 18.3, 1977, pp. 106–109.

[11] Opie, Keir, et al. Commodity-Specific Disaggregation of 2002 Freight Analysis Framework Data to County Level in New Jersey. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2121, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 128–134.

[12] Sachs, Jeffrey D., et al. *Macroeconomics in the Global Economy*. Prentice Hall, New Jersey, 2005.

[13] Sorratini, Jose A., Robert L. Smith, Jr.. Development of a Statewide Truck Trip Forecasting Model Based on Commodity Flows and Input-Output Coefficients. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1707, 2000, pp. 49–55.

[14] Viswanathan, Krishnan et al. Disaggregating Freight Analysis Framework Version 2 Data for Florida: Methodology and Results. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2049, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 167–175.

[15] Zavaterro, David A., and S. Weseman. Commercial Vehicle Trip Generation in the Chicago Region. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 834, 1981, pp. 12–15.