Fusing Freight Analysis Framework and Transearch Data: An Econometric Data Fusion Approach with Application to Florida

Salah Uddin Momtaz
PhD student
Department of Civil, Environmental & Construction Engineering
University of Central Florida, 12800 Pegasus Dr, Orlando, FL 32816.
Email: smomtaz@Knights.ucf.edu

Naveen Eluru
Associate Professor
Department of Civil, Environmental & Construction Engineering
University of Central Florida, 12800 Pegasus Dr, Orlando, FL 32816
Email: naveen.eluru@ucf.edu

Sabreena Anowar
Post-doctoral Researcher
Department of Civil, Environmental & Construction Engineering
University of Central Florida, 12800 Pegasus Dr, Orlando, FL 32816
Email: Sabreena.Anowar@ucf.edu

Nowreen Keya
PhD student
Department of Civil, Environmental & Construction Engineering
University of Central Florida, 12800 Pegasus Dr, Orlando, FL 32816
Email: nowreen.ce.buet@gmail.com

Bibhas Kumar Dey
PhD student
Department of Civil, Environmental & Construction Engineering
University of Central Florida, 12800 Pegasus Dr, Orlando, FL 32816
Email: bibhas.ce@Knights.ucf.edu

Abdul Pinjari
Associate Professor
Department of Civil Engineering
Indian Institute of Science (IISc), Bangalore, 560012. India
Email: abdul@iisc.ac.in

S. Frank Tabatabaee
Systems Transportation Modeler
Forecasting and Trends Office
Florida Department of Transportation, 605 Suwannee Street, Tallahassee, Florida 32399.
Email: Frank.Tabatabaee@dot.state.fl.us

ABSTRACT

A major hurdle in freight demand modeling has always been the lack of adequate data on freight movements for different industry sectors for planning applications. Both Freight Analysis Framework (FAF) and Transearch (TS) databases contain annualized commodity flow data. However, the representation of commodity flow in the two databases are inherently different. FAF flows represent estimated transportation network flows while TS flows represent production-consumption commodity flows. Our study aims to develop a fused database from FAF and TS to realize transportation network flows at a fine spatial resolution (county level) while accommodating for the production and consumption behavioral trends (provided by TS). Towards this end, we formulate and estimate a joint econometric model framework embedded within a network flow approach and grounded in maximum likelihood technique to estimate county level commodity flows. The algorithm is implemented for the commodity flow information from 2012 FAF data and 2011 TS databases to generate transportation network flows for 67 counties in Florida. The proposed approach can potentially circumvent the need for the purchase of expensive TS database for future years.

Keywords: Freight; FAF4; Transearch; Data Fusion; Fractional Split; Joint Model; Florida

INTRODUCTION

Study Motivation

On a daily basis, 122.5 million households, 7.5 million business establishments, and 90,000 governmental units in the United States rely heavily on the efficient movement of freight. In 2015, the country’s transportation system moved a daily average of about 49.3 million tons of freight valued at more than $52.5 billion. According to Bureau of Transportation Statistics (BTS), between 1998 and 2015, movement of freight (including imports and exports) grew by approximately 18 percent and is forecasted to increase by more than 40 percent by 2045 (Freight Facts and Figures, 2017). The increased volume of freight movements, within and across the country, may be attributed to rapid population and employment growth, economic expansion, continued globalisation, altering landscape of consumer and business preferences, and overwhelming popularity of e-commerce (Giuliano et al., 2018; Dablanc and Rodrigue, 2017). The growth in freight activities coupled with the continuing growth in passenger vehicle miles will undoubtedly put additional strain on the nation’s highway system in the form of additional congestion, traffic accidents, air pollution, noise, and expeditious deterioration of the highway surface.

Traditionally, the travel demand forecasting field has only focused on estimating passenger travel demand. Hedges (1971) suggested that one of the pre-requisites of a successful and effective urban transportation model is its ability to accommodate the interactions between passenger and freight trips. In recent years, there is growing recognition among travel demand modellers that freight planning is an important exercise for overall demand forecasting procedure. In the passenger travel realm, trip based or activity based demand models are often developed using household travel surveys (small or large scale) conducted in urban regions. Compared to passenger travel demand, a major hurdle in freight demand modeling has always been the lack of adequate, accessible, and reliable data on commodity movements amongst different industry sectors at a sufficiently fine geographic level for planning applications. The scarcity of a comprehensive freight flow dataset leaves the analyst with two choices: (1) develop a methodological tool based on available data, or (2) collect the necessary data for developing models either directly from shippers/carriers or indirectly through third party data providers (Giuliano et al., 2010). Extensive data collection approaches are costly and the validity of third party data is questionable. Moreover, despite the recent advances in freight travel demand modeling, the development of tools to estimate current and future freight flows has been limited. The current research effort is geared towards addressing the data availability challenge through an innovative econometric methodology for data fusion.

Comparison between FAF and Transearch Data

Several data sources are available for freight planning purposes in the United States. Of these, the most commonly adopted sources include Freight Analysis Framework (FAF), Transearch (TS), American Trucking Research Institute (ATRI) truck GPS data, and Department of Transportation (DOT) weigh-in-motion (WIM) data. FAF and TS databases contain annualized commodity flow data that can be used in long range freight forecasting. FAF database is a derivative product of the Commodity Flow Survey (CFS). CFS is a shipper based survey carried out every five years since 1997 as part of the Economic Census by the US Census Bureau, in partnership with Bureau of Transportation Statistics (BTS)). While FAF is derived from CFS data, several additional steps are undertaken to arrive at FAF flows. FAF supplements CFS by integrating variety of other sources for sectors not covered in the survey and estimating flows for commodities that are not included in CFS comprised of goods generated from imports (foreign establishments), publishing, farms, construction and demolition, logging services and fisheries. FAF data is freely available to the public and can be downloaded from the Federal Highway Administration (FHWA) website (FHWA and BTS, 2012). It provides freight flows (by weight, value and mode) for 43 commodity types classified by Standard Classification of Transported Goods (SCTG 2-digit) code. FAF geographic zones are determined based on the spatial resolution considered in the Commodity Flow Survey. The smallest spatial resolution at which CFS generates commodity flow estimates are the 132 domestic zones across the United States and 8 foreign zones (Hwang et al., 2016). The baseline year for current FAF data (FAF^4) is 2012 and includes forecasts on freight flows between 2015 and 2045 at a 5-year interval.

The Transearch database, a proprietary product developed by IHS Global Insight, provides detailed information on freight flows (by weight, value and mode) as well. The database is constructed from various commercial and public sources including: Annual Survey of Manufacturers (ASM), Surface Transportation Board (STB) Rail Waybill Sample, Army Corps of Engineers Waterborne Commerce data, Federal Aviation Administration (FAA), Enplanement Statistics, and Airport-to-airport cargo volumes. However, the algorithm used to generate the final data product is not publicly available. The freight flows in TS are reported by commodity type based on the Standard Transportation Commodity Code (STCC) in more than 500 categories. The data can be purchased at a fine spatial resolution (such as county level). However, the database is expensive to acquire and requires substantial investment from transportation agencies.

Although both FAF and TS provide annual commodity flows in the United States, several differences exist between these sources. The most obvious difference arises from the variability in data collection mechanism employed; FAF relies on processing commodity flow data (such as CFS 2012) while TS employs various sources of data to generate county level flows using a proprietary algorithm. A second difference arises from what the commodity flows in each dataset represent. FAF flows represent estimated transportation network flows while TS flows represent production-consumption commodity flows. To illustrate the difference, consider that X units of a commodity is shipped from location A (production zone) to location B (consumption zone) through an intermediate location C. The FAF flows would represent these flows as X units from A to C and X units form C to B. On the other hand, in TS, these flows are only represented as X units from A to B. Thus, FAF would report a total tonnage of 2X units transferred while TS would report only a transfer of X units. A more general summary of data sampling procedures in FAF and TS is presented in Figure 1. From the figure, it is evident that FAF flows are potentially sampled at more intermediate points such as warehousing locations while TS flows are considered only at origin and destination.

For understanding transportation network usage measured through network flows, FAF is a more appropriate database as the reporting is based on realized network flows. On the other hand, the flows represented in the TS database are annual production-consumption measures from the TS defined regions and do not represent the estimated transportation network path flows. To be sure, there is significant value in understanding production and consumption trends to develop a behavioral framework of freight commodity flows in the future. In terms of cost, FAF data is freely available while TS database is an expensive database and the algorithm employed is inaccessible to users. The commodity type definition across the two datasets is also different – 43 commodity types in FAF and over 500 commodity types in TS. Finally, the coarser spatial and commodity type resolution in FAF makes it challenging to generate reliable network flow estimates. While TS provides data at a fine spatial and commodity type resolution, the production consumption behavior of the database requires additional analysis to realize transportation network flows. Overall, the comparison highlights the inherent strengths and weaknesses of the two databases.

Current Study Context

FAF is a comprehensive database providing useful information for evaluating the impact of freight movement on transportation network (Hwang et al., 2016). However, it has limited use for freight planning and decision-making processes at the state, district, or Metropolitan Planning Organizations (MPO) due to its high level of spatial aggregation (Roman-Rodriguez et al., 2014; Harris et al., 2010). Thus, several research efforts have attempted to address this spatial resolution challenge with FAF data. A summary of earlier studies that attempted to merge or disaggregate different freight data sources is provided in Table 1. The table provides information on the study area, datasets employed, objective(s) of the research effort, modeling methodology employed, and exogenous variables considered.

Several observations can be made from the table. First, the primary objective of majority of the studies is on developing a procedure for disaggregating FAF data from the FAF zone level to a county level or traffic analysis zone (TAZ) level. Second, the states in the US which have developed disaggregation procedures include Texas, California, New Jersey, Wisconsin, Georgia, and Florida. Third, the various methods considered to disaggregate FAF flows include: (i) proportional weighting method, and (ii) statistical methods. In the proportional weighting method, a “disaggregation factor” is estimated using various socio-economic variables (such as employment and population), land use (occupied by ports), truck flows, and truck VMT variables by computing the ratio of the variables of interest at the disaggregate spatial resolution and aggregate spatial resolution. Using these factors, the freight flows are allocated to the disaggregate spatial resolution. The disaggregation factors are considered to vary based on the type of origin and destination spatial configuration (such as internal – internal zonal pair, external – internal zonal pair). The statistical methods considered in freight modeling include linear or log-linear regression, structural equation modeling, economic input output models, and fractional split methods that employ socio-economic and demographic variables such as employment and population as exogenous variables. The models developed are employed to generate freight flows at the desired disaggregate spatial resolution. These models are typically validated by aggregating freight flows at the finer resolution and comparing it to the observed flows at the aggregate resolution. Fourth, in the disaggregation studies, the variables of interest include tonnage, value and/or ton-miles. Finally, the variables considered to be of significance in the data merging process are: employment, population, travel time and cost, business establishments, and transportation system characteristics.

Based on the literature review, it is evident that multiple research efforts have attempted disaggregation of FAF commodity flow to a lower spatial resolution such as county or TAZ. While the disaggregation is of immense value, the approach employed is purely a factoring exercise without any attempt to address production consumption relationships. FAF data inherently does not provide production consumption relationship and hence, using FAF alone to arrive at production consumption flows is not possible. To be sure, several earlier research studies employed TS flows for validating FAF disaggregation outputs (Opie et al., 2009; Ruan & Lin, 2010; Viswanathan et al., 2008). In our study, we enhance earlier research efforts by developing a fusion framework that disaggregates FAF flows while accounting for production consumption relationships observed in TS.

In summary, the primary motivation for our study is the development of a fused database to realize transportation network flows at a fine spatial resolution (county level) while accommodating for production and consumption behavioral trends. Thus, we undertake disaggregation of FAF flows while augmenting with production consumption based TS flows. Towards this end, we formulate and estimate a joint econometric model framework embedded within a network flow approach grounded in maximum likelihood technique to estimate county level commodity flows. The framework has two separate modules to ensure matching estimated county level flows with commodity flows in FAF and TS at the appropriate spatial resolution. A third module generates a behavioral connection between FAF and TS. In our algorithm, we connect the flows between TS and FAF by generating potential paths between the origin and destination of interest for TS flows. Note that the inherent differences in the data cannot be completely reconciled. Hence, the framework focuses on building a fused database that maximizes the match with the commodity flows in the two databases. The consideration of behavioral trends in the model framework can assist us in parameterizing TS flow relationships thus allowing us to circumvent TS for the future (if needed). The proposed algorithm is implemented for the commodity flow information from 2012 FAF data for five FAF zones and 2011 TS databases for 67 counties in Florida.

The remainder of the paper is organized as follows. Mathematical formulation details are presented in the econometric model framework section. The empirical data and data preparation steps are discussed in empirical data section. The fifth section presents the results of the proposed data fusion algorithm along with validation of outputs. The final section concludes the paper with a discussion of the limitations of the current study and directions for future research.

ECONOMETRIC FRAMEWORK

In this section, the proposed algorithm is described. Prior to discussing the algorithm details, the notations and terminology used in the algorithm are presented.

Network Representation

The study defines nodes, paths, and links in the usual network theoretic approach. Nodes represent county centroids. These represent either origin, destination or intermediate points. A direct connection between any two nodes is defined as a link. Paths represent a series of links that connect an origin and destination. To elaborate on the terminology, a simple representation is provided in Figure 2. In Figure 2(a), from origin county ‘A’, freight flow can be transferred to destination county ‘B’ via a direct path (i.e. no intermediate nodes) which is indicated by a solid line. The flow could also move along an indirect path. In our study, given the model is a statewide model, we assume that one intermediate node is adequate for considering all possible paths between OD pairs to ensure computational tractability of the algorithm. The path with one intermediate node is referred to as one-hop path. In Figure 2(a), a one-hop path from county ‘A’ to county ‘B’ with an intermediate stop at county ‘C’ is shown with the dashed line. In Figure 2(b), origin node ‘1’ and destination node ‘4’ have the following possible paths on the network. (i) ‘1’ - ‘4’ direct path (link ‘1’ – say, path 1), (ii) ‘1’ - ‘3’ - ‘4’ is a one-hop path (link ‘2’ – link ‘3’ – say path 2, or link ‘2’ – link ‘6’ – say path 3). Therefore, three different paths are considered here from origin ‘1’ to destination ‘4’ that uses four different links (i.e. links ‘1’, ‘2’, ‘3’, and ‘6’).

To represent the relationship between paths and links in our system, a link path matrix is generated. For the network in Figure 2(a) and 2(b), the link-path matrix (A) is shown in Figure 2(c). The rows represent the links and the columns represent the paths between the given OD pairs (see Figure 2 for details). Each element of the matrix is a binary indicator that represents if the link ‘i’ is included in the corresponding path. The variable of interest in the algorithm is the transportation network county to county flows generated by fusing TS data at the county level and FAF data at the FAF region level. Let V_{ij} represent the link flows between county pair i and j. The entire set of link flows are considered in a matrix form as V. Given the link-path matrix A, and path flow vector ‘h’, the link flow matrix, ‘V’ is given by the following equation.

$$V = A * h \tag{1}$$

Joint Model System

Let, y_{ij} represent the natural logarithm of the reported TS flow, and $\hat{y}_{ij}$ the estimated transearch flow. With these notations, the log-linear model takes the following form:

$$\hat{y}_{ij} = \ \beta X_{ij} \tag{2}$$

where, X_{ij} are the independent variables for the specific county pair $i - j$ and β represents the corresponding vector of parameters. Assuming the usual linear regression formulation, the likelihood for the estimation takes the following form:

$$LL_{TS_{i,j}} = \frac{\emptyset(\frac{\hat{y}_{ij} - y_{ij}}{\sigma_{TS}})}{\sigma_{TS}} \tag{3}$$

where, $\emptyset$ represent the probability density function of the standard normal distribution, and σ_{TS} is the standard deviation of ε_{ij}.

Given that TS flow is an input-output flow, the objective is to decompose these flows into estimated network level link flows by considering the various paths between each OD pair. The path flows will allow us to determine the link flows. These flows are generated by employing a fractional split approach. The actual path flow is unobserved; hence, a latent variable is considered and the resulting link flows are matched with the observed flows. The probability for each path is determined using a random utility approach. The proposed approach recognizes that the proportion of the commodity flow assigned to a path is influenced by the utility of the path relative to other alternative paths. The mathematical formulation employed is as follows:

$$\cup_{ij}^{k} = \sum_{i,j \in O,D;\, k=1}^{K} \alpha\, X_{ij}^{k} \tag{4}$$

$$P\left(k_{ij} \middle| x_{ij}^{k}\right) = \frac{\exp(\cup_{ij}^{k})}{\sum_{l=1}^{K} \exp(\cup_{ij}^{l})} \tag{5}$$

$\cup_{ij}^{k}$ represents the utility for the k^{th} path between i and j; α represents the vector of parameters for path utility and $P\left(k_{ij} \middle| x_{ij}^{k}\right)$ $r$espresents the probability for the k^{th} path between i and j. Based on the path flow probability the flow assigned to each path is determined as follows:

$$h_{ij}^{k} = \hat{y}_{ij} * P(k_{ij}|x_{ij}^{k}) \tag{6}$$

The path flow estimation leads to the estimation of the link flows V, using Equation (1). Given that these flows are available at the county level, we need to aggregate them to a coarser level to compare the flows to observed FAF flows. The aggregation is achieved over Origin (*O*) and Destination (*D*) FAF zone as:

$$\hat{F}_{OD} = \sum\nolimits_{i \in O, j \in D} V_{ij} \qquad \forall\, O, D \in \Theta \tag{7}$$

where i, j represent counties in *O* and *D* respectively and V_{ij} represents the corresponding link flow between county i and county j; where Θ is set of all FAF zones. Let F_{OD} be the observed FAF flows. The log-likelihood for comparing the predicted FAF flows with observed FAF flows in the linear regression form is given by the following mathematical expression, where, σ_{FAF} is the standard deviation of the estimate of FAF flows.

$$LL_{FAF} = \frac{\emptyset(\frac{\widehat{F}_{OD} - F_{OD}}{\sigma_{FAF}})}{\sigma_{FAF}} \quad (8)$$

Given the aggregation proposed, the contribution of the FAF log-likelihood needs to be carefully computed. While origin and destination counties have their corresponding FAF zones, the intermediate zones also have a FAF zone. Therefore, the allocation is obtained for an OD pair by apportioning the error to all FAF zones involved over the entire path set for that OD pair. For this purpose:

$$LL_{FAF}{}^{k_{ij}} = \frac{\sum_{r=1}^{n} LL_{FAF}{}^{r}}{n} \quad (9)$$

where, n is the number of link in the path $k = \begin{cases} 1, & for\ direct\ path \\ 2, for\ one - hop\ paths \end{cases}$

Further, FAF zones can represent a large number of counties. To normalize for the number of counties, we employ the following equation:

$$LL_{FAF}{}^{OD,Norm}{}_{i,j} = \frac{\sum_{s=1}^{N} LL_{FAF}{}^{k_{ij}}}{N_C} \quad (10)$$

where, N_c is the number of county pairs in the OD FAF region pairs. Finally, the joint log-likelihood is provided by the sum of log-likelihood for FAF and TS flow.

$$LL_{total\ i,j} = \sum_{i,\ j} (\ln(LL_{TS_{i,j}}) + \ln(LL_{FAF}{}^{OD,Norm}{}_{i,j})) \quad (11)$$

A flowchart describing the econometric modeling approach is provided in Figure 3. The proposed algorithm is programmed in Gauss matrix programming language (Aptech, 2015).

DATA PREPARATION

In this section, we briefly discuss the data preparation steps. Florida has five FAF regions: Jacksonville, Miami, Orlando, Tampa, and remainder of Florida (see Figure 4). On the other hand, the state is represented as 68 zones in the TS database. In our study, we have access to the 2011 base year data for Florida that includes forecasts for 2015 through 2040 at a five-year interval.

Commodity Classification

As mentioned before, there are 43 commodity types in FAF while TS commodities are classified into 562 commodity types. To generate a comparable commodity type classification, we consolidated the different commodity types in the two databases into 13 commodity types (see, Viswanathan et al., 2008 for a similar classification of commodities). The consolidated commodity types are: (1) agricultural products, (2) minerals, (3) coal, (4) food, (5) nondurable manufacturing, (6) lumber, (7) chemicals, (8) paper, (9) petroleum, (10) other durable manufacturing, (11) clay and stone, (11) waste, (12) miscellaneous freight (including warehousing) and (13) unknown. Table 2 provides a comparison of freight flows by the consolidated commodity types within Florida. The highest variation in flow is observed for non-durable manufacturing and chemicals commodity type (6). The lowest ratio is observed for miscellaneous freight and warehousing commodity type (0.28). Note that TS reports secondary flows including drayage whereas FAF does not contain any information on drayage. Thus, it is not surprising that we have a lower ratio.

Independent Variables Generation

We compiled several exogenous variables for the fusion model. These are: (1) origin-destination indicator variables including origin (or destination) is in Orlando, Tampa, Jacksonville, Miami, Remainder of Florida region, (2) socio-demographic and socio-economic indicators including population and employment, (3) transportation infrastructure indicators including road and railway line length, number of ports, airports, and intermodal facilities, and (4) several interactions of these variables. Population and employment data were collected at the county level from the U.S. Census Bureau (U.S.C. Bureau, 2017a, 2017b). Transportation related variables were generated using the ArcGIS platform intersecting the facility shapefiles collected from Florida Geographic Data Library (FGDL, 2017) with that of the county shapefile. Post-processing of the intersected files provided us the length of roadways and railways, number of seaports, airports, and intermodal facilities at the county level. Please note that these variables were compiled for the base year of 2011.

Finally, for the fractional split model, we needed to generate all path choice set for every OD pair. For this purpose, we considered 1 direct path and 66 one-hop paths (that pass through another county). The paths were generated for all OD pairs with non-zero flow. The overall path matrix was quite large with number of elements ranging from 6700 to 270000 across various commodities. For the paths created, path distances between origin and destination counties were generated as a sum of the link distances. A link distance for county pairs was determined using the shortest path procedure of ArcGIS's network OD cost tool. The highway route for the local and highways provided by the Florida Department of Transportation (FDOT) was used for this purpose.

IMPLEMENTATION OF DATA FUSION ALGORITHM

The proposed algorithm is implemented separately for each commodity type. For the sake of brevity, we only present the model results for two commodities: Agricultural products and Food. The results for the other commodities are available from the authors upon request. We discuss the results for the two commodities separately.

Commodity Type: Agricultural Products

In Table 3, columns 3 and 4 provide parameter estimates and t-statistics for Agricultural products. The TS module corresponds to the overall county to county flow tonnage while the FAF module provides the fractional split model estimates.

TS Module

In terms of Origin indicator variables, for agricultural products, Jacksonville origin region is likely to have lower flow relative to other locations. On the other hand, Miami origin is associated with larger flows. For Destination indicator variables, Orlando is associated with larger flows while Miami is associated with smaller flows. The overall regional trends closely align with the trends of agricultural commodity generation reported by Hodges and Rahmani, 2008. The reader would note that these indicator variables serve as region specific constants and are influenced by other exogenous variables. To elaborate, these variables represent the inherent influence of regions specific characteristics not considered in the exogenous variables.

For agricultural products, several destination specific attributes have significant impact on flows. The number of warehouses in the destination county is associated positively with flows to the destination county. The number of intermodal facilities in the destination county is negatively associated with flows. The reader would note that while the impact of intermodal facilities appears to be negative on first glance, it needs to be recognized that increased inter-modal facilities in a county is also likely to have a higher number of warehouses. Thus, the model result for these parameters need to be considered together. On the other hand, no attributes for the origin location provided significant parameters. Several interaction variables from different variable categories were also considered. The interaction term between origin county employment and destination county employment was found to be positively associated with county to county flows. Specifically, a higher rate of employment at the origin county relative to employment at the destination county indicates higher movement of agricultural product commodity. The standard error of the estimate represents the standard deviation of the unobserved component in the regression model.

FAF Module

The fractional split model in the FAF module is based on a large number of alternatives. Hence, the model only allows for the estimation of generic coefficients i.e. no alternative specific effects can be estimated. The path distance variable is considered in the model. Any other origin or destination variable would require us to consider interaction with path distance. The models with such interaction variables did not provide intuitive results. Hence, we resorted to considering only the path distance variable in our FAF module. The path distance variable was negative as expected, indicating that longer distance reduces the probability of the paths being chosen. The result clearly indicates a larger path flow allocation to direct paths and a smaller flow allocation to one-hop paths.

Commodity Type: Food

In Table 3, columns 5 and 6 provide parameter estimates and t-statistics for Food.

TS Module

For Food commodity, indicator variables for Tampa and Orlando origins are positively associated with flows. The magnitude of coefficient for the Tampa region is larger than the corresponding magnitude of coefficient for Orlando region. In terms of destination county attributes, number of ports and road network length in the destination counties are associated positively with food flows. The flow is also influenced by origin county road network length.

The interaction between origin county employment and destination county population was negatively associated with Food flows. The result might be indicating that an increase in origin county employment reduces the need for transporting Food products further from the origin. The parameter impacts while intuitive in general also necessitate the need for further analysis using data from different spatial and temporal regions.

FAF Module

Similar to the model for agricultural products, we found negative relationship between the path distance and the path flow proportions in the model for food as well. The magnitude of the parameter is substantially larger for Food relative to Agricultural products. To be sure, these two parameters are not directly comparable.

Model Validation

To evaluate the performance of our proposed algorithm, several validation exercises were conducted. To be sure, the county to county transportation flows generated from the exercise do not have an observed counterpart to validate. Hence, we resort to validation by examining the outputs. After fusing FAF and TS databases, we compare the transportation link flows obtained with the production consumption flows. For example, the ratio of FAF and TS for agricultural products is 2 (see Table 2). The ratio of the fused flows with TS flows was found to be 1.45 (see Table 4). The reader would note that while Transearch flows are non-zero for only a subset of county flows, the transportation network flows are theoretically non-zero for all pairs. Hence, the number of observations in Table 4 for fused flows are (67*67=4489). A similar exercise for Food yielded a value of 1.62 (relative to the original ratio of 2.40). In both cases, the results are quite reasonable (see Table 4). To further characterize these differences, we compare flows originating (or destined) from (to) a county for Transearch flows and fused flows. The comparison is undertaken by computing the percentage of total flows originating (or destined) from (to) each county. Then the percentage point difference for each county for Transearch and fused flows is computed. The results from the comparison are presented in Table 5. We present the mean and standard deviation of the differences over all the counties in our analysis by origin and destination county. The fused flows for agricultural products commodity show larger variation from Transearch flows – 2.30 points for origin counties and 2.29 points for destination counties. On the other hand for food commodity, the difference between fused flows and Transearch flows is within a narrower range – 1.37 points for origin counties and 1.44 points for destination counties. The standard deviation results also confirm the trend – larger variability for agricultural products commodity. While these results are quite informative, it is important to recognize that the differences do not necessarily reflect error in the fused flows.

As a second step, we plot the relationship between county to county flows for TS and fused flows. The plots are created by considering proportion of statewide flows originating (or destined) to each county. Figures 5 and 6 provides the plots for Agricultural Products and Food, respectively. In these figures, the plots for TS are on the left and the plots for fused flows are on the right. We can see from the figures that for Agricultural Products, both origin and destination based plots, are quite similar. The counties in Central and South Florida regions account for larger share of the flows in TS as well as fused flows. For Food, the fused flows indicate a larger share of flows in Central and South Florida relative to TS flows. However, the overall trends are still very similar.

As a final comparison exercise, we compare TS and fused flows originating from Miami-Dade County for the two commodities. For this purpose, we plot the tonnage of flows transferred between counties (see Figure 7). For TS, the reported flows originating form Miami-Dade to all counties are plotted as direct flows on the network for illustration. For fused flows, the path flows estimated from our algorithm for Miami-Dade to all counties are plotted. The thicker the line on the road network, the larger is the tonnage transferred. From the figure, it is evident that we observe substantially thicker lines for fused flows. This is expected because fused flows should represent network flows whereas TS flows only represent origin destination flows. Hence, they always are likely to pass directly, whereas fused flows would be a result for multiple origin destination flows. Overall, the three validation steps provide evidence that the fusion algorithm provides outputs as expected from a joint system disaggregating FAF with production consumption trends form TS.

CONCLUSIONS

A major hurdle in freight demand modeling has always been lack of adequate data on commodity movements amongst different industry sectors for planning applications. Several data sources are available for freight planning purpose in the United States. Of these, the two most commonly adopted sources are Freight Analysis Framework (FAF) and Transearch (TS). FAF (freely available) and TS (proprietary) databases contain annualized commodity flow data that can be used in long range freight forecasting. Although both FAF and Transearch provide annual commodity flows in the United States, several differences exist between these sources, including the variability in data collection mechanism employed, and variability in the spatial and commodity type resolution. The coarser spatial resolution in FAF makes it challenging to generate reliable network flow estimates. While TS provides data at a fine spatial resolution, the supply demand nature of the database does not represent the actual transportation network path flows and requires additional analysis to realize transportation network flows. The primary motivation for our study is the development of a fused database to realize transportation network flows at a fine spatial resolution (county level) while accommodating for production and consumption behavioral trends. Clearly, the level of detail provided by FAF data would be much enhanced through the disaggregation of this data from a zonal level to a county level. The disaggregated commodity flow data will allow for the estimation of truck trips, which will be useful for both regional and local level freight demand forecasting.

To achieve the goal of the study, we undertake disaggregation of FAF flows while augmenting with production consumption based TS flows. Towards this end, we formulate and estimate a joint econometric model framework embedded within a network flow approach grounded in maximum likelihood technique to estimate county level commodity flows. The algorithm is implemented for the commodity flow information from 2012 FAF data for five FAF zones and 2011 TS databases for 67 counties in Florida. Overall, our model system predicted well as manifested from the ratio of fused flows to observed TS flows for the two commodities for which the results are presented (Agricultural Products and Food). Moreover, the path distance coefficients are intuitive. As expected, shorter paths are allocated higher fraction of the flows compared to the longer paths. The fusion algorithm can be applied to obtain fused flows for future years without having to purchase expensive TS dataset.

To be sure, the study is not without limitations. In our algorithm, only one-hop paths are considered for computational tractability. It would be interesting to examine how the fused outputs are influenced by a larger choice set of paths. It might also be interesting to examine the spatial and temporal transferability of the proposed algorithm using either past or future data.

ACKNOWLEDGMENT

This work is funded by Florida Department of Transportation (FDOT) and has been carried out at the University of Central Florida. The content of this paper reflects the views of the authors who are solely responsible for the facts, opinions, findings and conclusions presented herein. The contents do not necessarily reflect the official views or policies of the Florida DOT.

REFERENCES

Aly, S., & Regan, A. (2014). Disaggregating FAF 2 Data For California. *Journal of Behavioural Economics, Finance, Entrepreneurship, Accounting and Transport*, *2*(2), 47–57.

Aptech. (2015). *Home - Aptech* (Version 14.0). Retrieved from https://www.aptech.com/

Bujanda, A., Villa, J., & Williams, J. (2014). Development of Statewide Freight Flows Assignment Using the Freight Analysis Framework (Faf3). *Journal of Behavioural Economics, Finance, Entrepreneurship, Accounting and Transport*, *2*(2), 47–57. https://doi.org/10.12691/JBE-2-2-3

Dablanc, L., & Rodrigue, J.-P. (2017). The Geography of Urban Freight. *The Geography of Urban Transportation*, *34*.

Federal Highway Administration (FHWA), & Bureau of Transportation Statistics (BTS). (2012). Freight Analysis Framework (FAF). Retrieved July 30, 2017, from http://faf.ornl.gov/fafweb/Default.aspx

Giuliano, G., Gordon, P., Pan, Q., Park, J., & Wang, L. (2010). Estimating Freight Flows for Metropolitan Area Highway Networks Using Secondary Data Sources. *Networks and Spatial Economics*, *10*(1), 73–91. https://doi.org/10.1007/s11067-007-9024-9

Giuliano, G., Kang, S., & Yuan, Q. (2018). Using proxies to describe the metropolitan freight landscape. *Urban Studies*, *55*(6), 1346–1363. https://doi.org/10.1177/0042098017691438

Harris, G. A., Anderson, M. D., Farrington, P. A., Schoening, N. C., Swain, J. J., & Sharma, N. S. (2010). Developing Freight Analysis Zones at a State Level: A Cluster Analysis Approach. *Journal of the Transportation Research Forum*, *49*(1), 59–68. https://doi.org/10.5399/osu/jtrf.49.1.2521

Hedges, C. (1971). Demand Forecasting and Development of a Framework for Analysis of

Urban Commodity Flow: Statement of the Problem. *Special Report 120: Urban Commodity Flow*, 145–148.

Hodges, A. W., & Rahmani, M. (2008). Economic Contributions of Florida's Agricultural, Natural Resource, Food and Kindred Product Manufacturing and Distribution, and Service Industries in 2008. *EDIS document FE829,* Food and Resource Economics Department, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL.

Hwang, H.L., Hargrove, S., Chin, S.M., Wilson, D.W., Lim, H., Chen, J., Taylor, R., Peterson, B. and Davidson, D., 2016. The Freight Analysis Framework Verson 4 (FAF4)-Building the FAF4 Regional Database: Data Sources and Estimation Methodologies (No. ORNL/TM-2016/489). Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States).

Lim, R., Qian, Z. (Sean), & Zhang, H. M. (2014). Development of a Freight Demand Model with an Application to California. *International Journal of Transportation Science and Technology*, *3*(1), 19–38. https://doi.org/10.1260/2046-0430.3.1.19

Mitra, S., & Tolliver, D. (2009). Framework for Modeling Statewide Freight Movement Using Publicly Available Data. *Journal of the Transportation Research Forum*, *48*(2), 83–102. https://doi.org/10.5399/osu/jtrf.48.2.2281

Opie, K., Rowinski, J., & Spasovic, L. (2009). Commodity-Specific Disaggregation of 2002 Freight Analysis Framework Data to County Level in New Jersey. *Transportation Research Record: Journal of the Transportation Research Board*, *2121*, 128–134. https://doi.org/10.3141/2121-14

Ranaiefar, F., Chow, J. Y., Rodriguez-Roman, Daniel Camargo, P. V, & Ritchie, S. G. (2013). UCI-ITS-WP-12-4 Geographic Scalability and Supply Chain Elasticity of a Structural Commodity Generation Model Using Public Data. In *92nd TRB Annual*

Meeting. Washinton D.C.

Rodriguez-Roman, D., Masoud, N., Jeong, K., & Ritchie, S. (2014). Goal Programming Approach to Allocate Freight Analysis Framework Mode Flow Data. *Transportation Research Record: Journal of the Transportation Research Board*, *2411*, 82–89. https://doi.org/10.3141/2411-10

Ross, C., Kumar, A., Wang, F., & Hylton, P. (2016). *Georgia DOT Research Project 13-27 Final Report Freight Movement, Port Facilities, and Economic Competitiveness – Supplemental Task: County-to- County Freight Movement (National and State Level)*. Atlanta, GA: Georgia Tech Research Corporation.

Ruan, M., & Lin, J. (2010). Synthesis Framework for Generating County-Level Freight Data Using Public Sources for Spatial Autocorrelation Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *2160*, 151–161. https://doi.org/10.3141/2160-16

Sorratini, J., & Smith, R. (2000). Development of a Statewide Truck Trip Forecasting Model Based on Commodity Flows and Input-Output Coefficients. *Transportation Research Record: Journal of the Transportation Research Board*, *1707*, 49–55. https://doi.org/10.3141/1707-06

Sprung, M. (2018). *Freight Facts and Figures 2017*. United States. Bureau of Transportation Statistics. https://doi.org/10.21949/1501488

The Florida Geographic Data Library (FGDL). (2017). FGDL Search/ Download Data. Retrieved July 31, 2017, from http://www.fgdl.org/metadataexplorer/explorer.jsp

U.S. Census Bureau. (n.d.). American FactFinder. Retrieved July 30, 2017, from https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

US Census Bureau. (2017). Population Clock. Retrieved July 30, 2017, from https://www.census.gov/popclock/

Viswanathan, K., Beagan, D., Mysore, V., & Srinivasan, N. (2008). Disaggregating Freight Analysis Framework Version 2 Data for Florida: Methodology and Results. *Transportation Research Record: Journal of the Transportation Research Board*, *2049*, 167–175. https://doi.org/10.3141/2049-20

FIGURE 1 FAF and Transearch Data Collection Methods and Dataset Generation

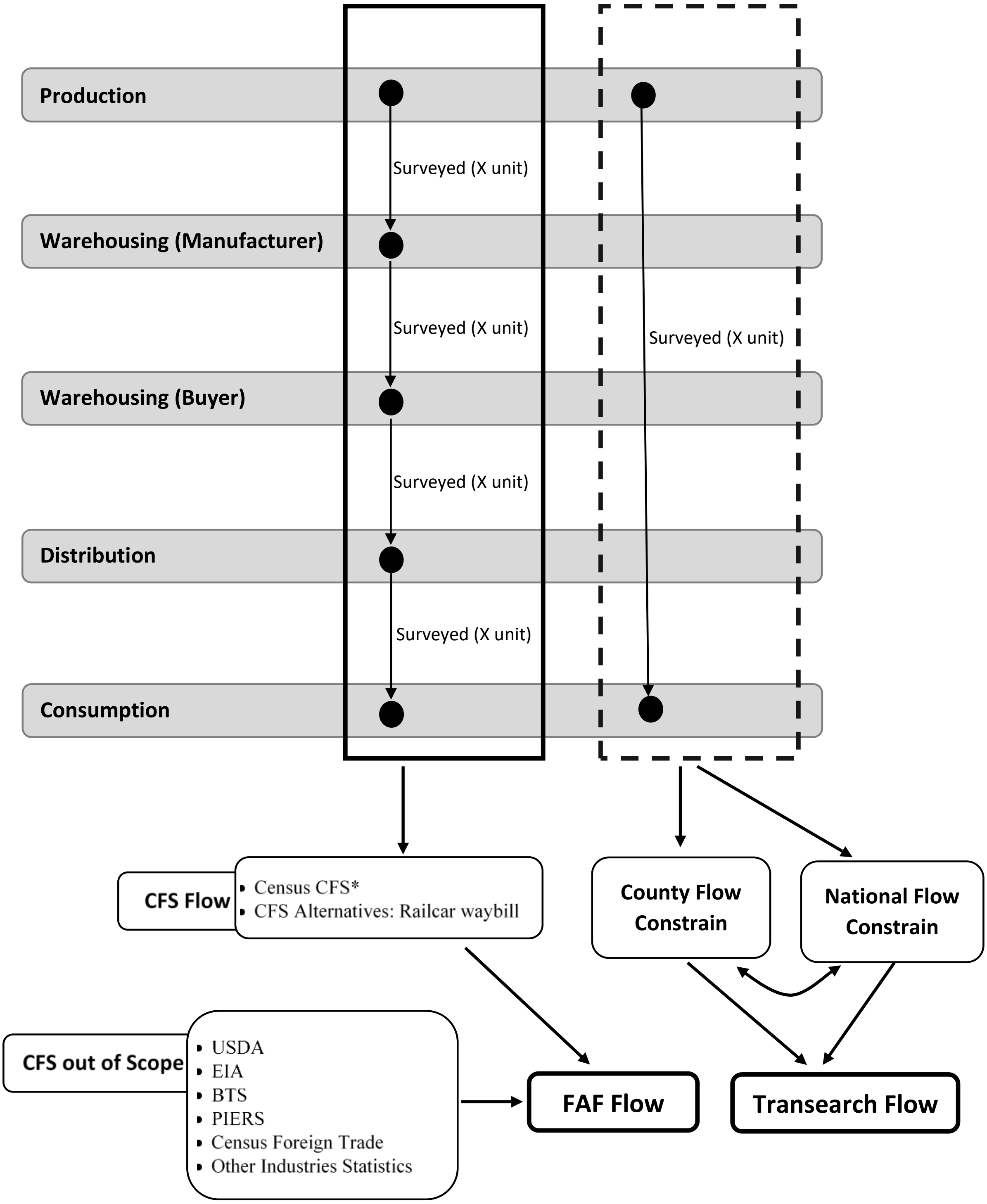
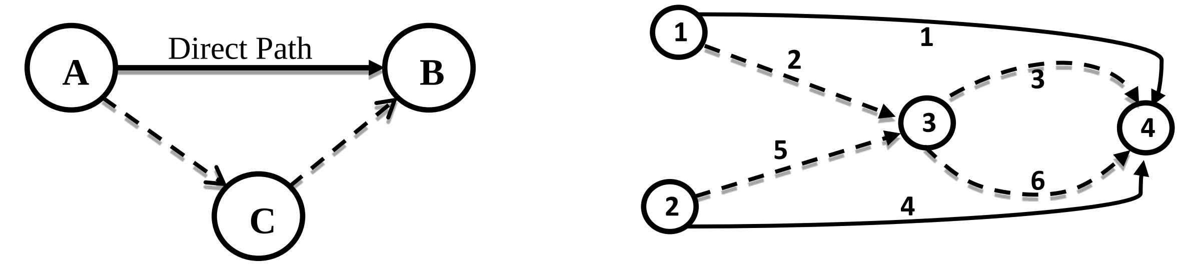


FIGURE 2 Paths, Links, and Nodes of a Simple Transportation Network



(a) Paths between OD pairs A and B

(b) Links and nodes on a network

$$A = \begin{array}{c|cccccc} & \multicolumn{3}{c}{O-D} & \multicolumn{3}{c}{O-D} \\ & \multicolumn{3}{c}{1-4} & \multicolumn{3}{c}{2-4} \\ Link\backslash Path & 1 & 2 & 3 & 1 & 2 & 3 \\ \hline 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 1 & 0 \\ 4 & 0 & 0 & 0 & 1 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 1 & 1 \\ 6 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}$$

(c) Links- path matrix for the simple network shown on (b)

FIGURE 3 Econometric Modeling Approach

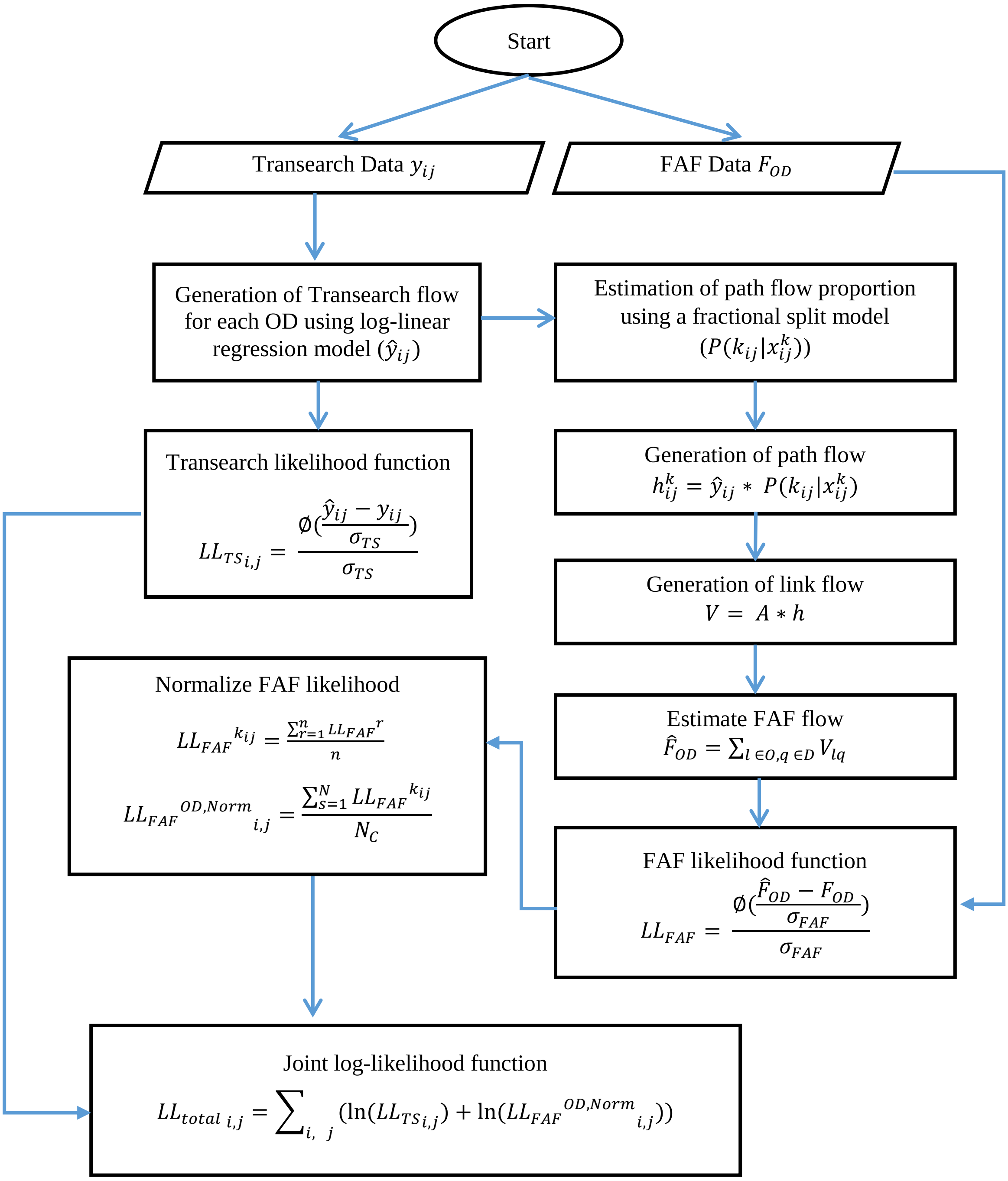

Start
Transearch Data y_{ij}
FAF Data F_{OD}
Generation of Transearch flow for each OD using log-linear regression model ($\hat{y}_{ij}$)
Estimation of path flow proportion using a fractional split model ($P(k_{ij}|x_{ij}^{k})$)
Transearch likelihood function
$LL_{TS_{i,j}} = \frac{\emptyset(\frac{\hat{y}_{ij} - y_{ij}}{\sigma_{TS}})}{\sigma_{TS}}$
Generation of path flow
$h_{ij}^{k} = \hat{y}_{ij} * P(k_{ij}|x_{ij}^{k})$
Generation of link flow
$V = A * h$
Normalize FAF likelihood
$LL_{FAF}{}^{k_{ij}} = \frac{\sum_{r=1}^{n} LL_{FAF}{}^{r}}{n}$
$LL_{FAF}{}^{OD,Norm}{}_{i,j} = \frac{\sum_{s=1}^{N} LL_{FAF}{}^{k_{ij}}}{N_C}$
Estimate FAF flow
$\hat{F}_{OD} = \sum_{l \in O, q \in D} V_{lq}$
FAF likelihood function
$LL_{FAF} = \frac{\emptyset(\frac{\hat{F}_{OD} - F_{OD}}{\sigma_{FAF}})}{\sigma_{FAF}}$
Joint log-likelihood function
$LL_{total\ i,j} = \sum_{i,\ j} (\ln(LL_{TS_{i,j}}) + \ln(LL_{FAF}{}^{OD,Norm}{}_{i,j}))$

FIGURE 4 FAF and Transearch TAZ

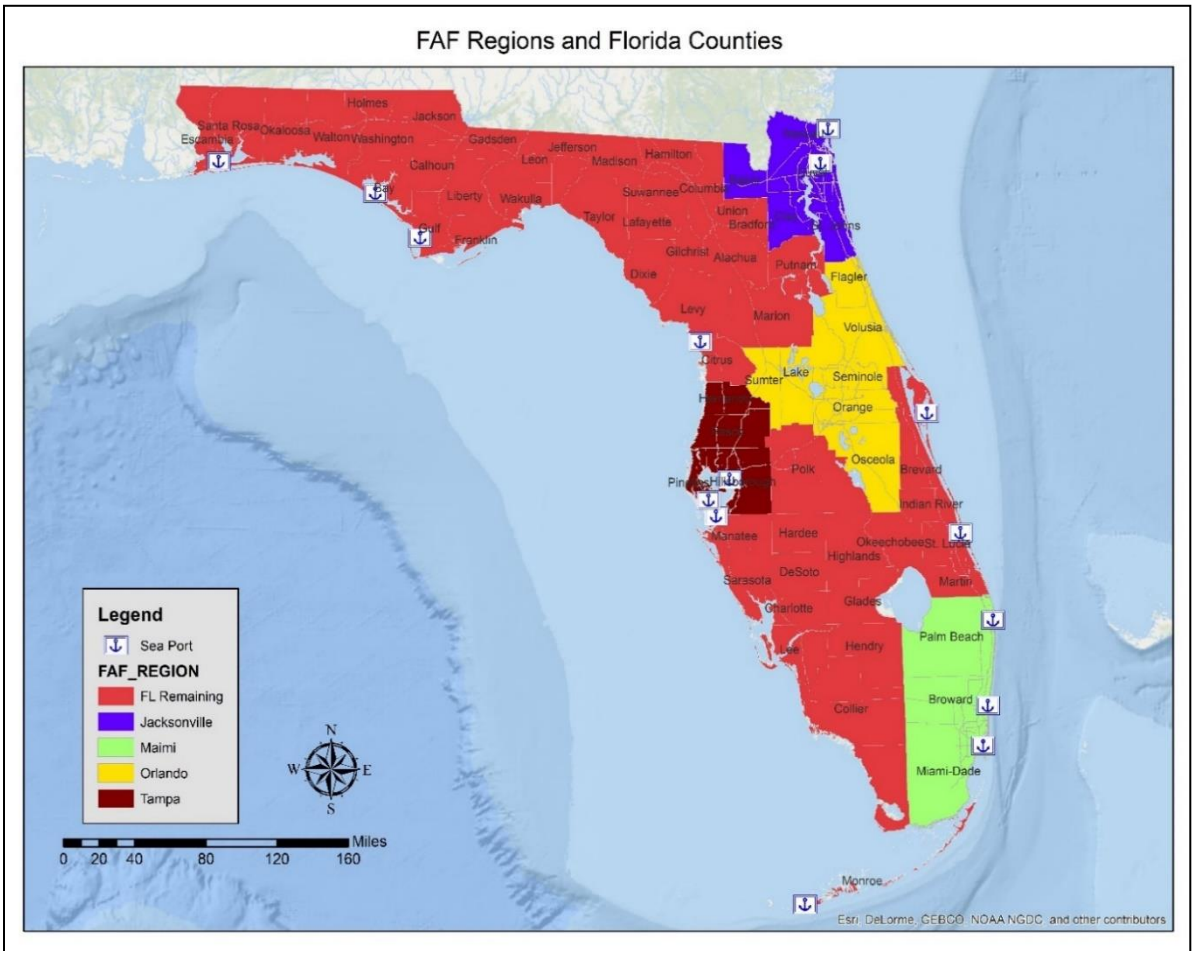
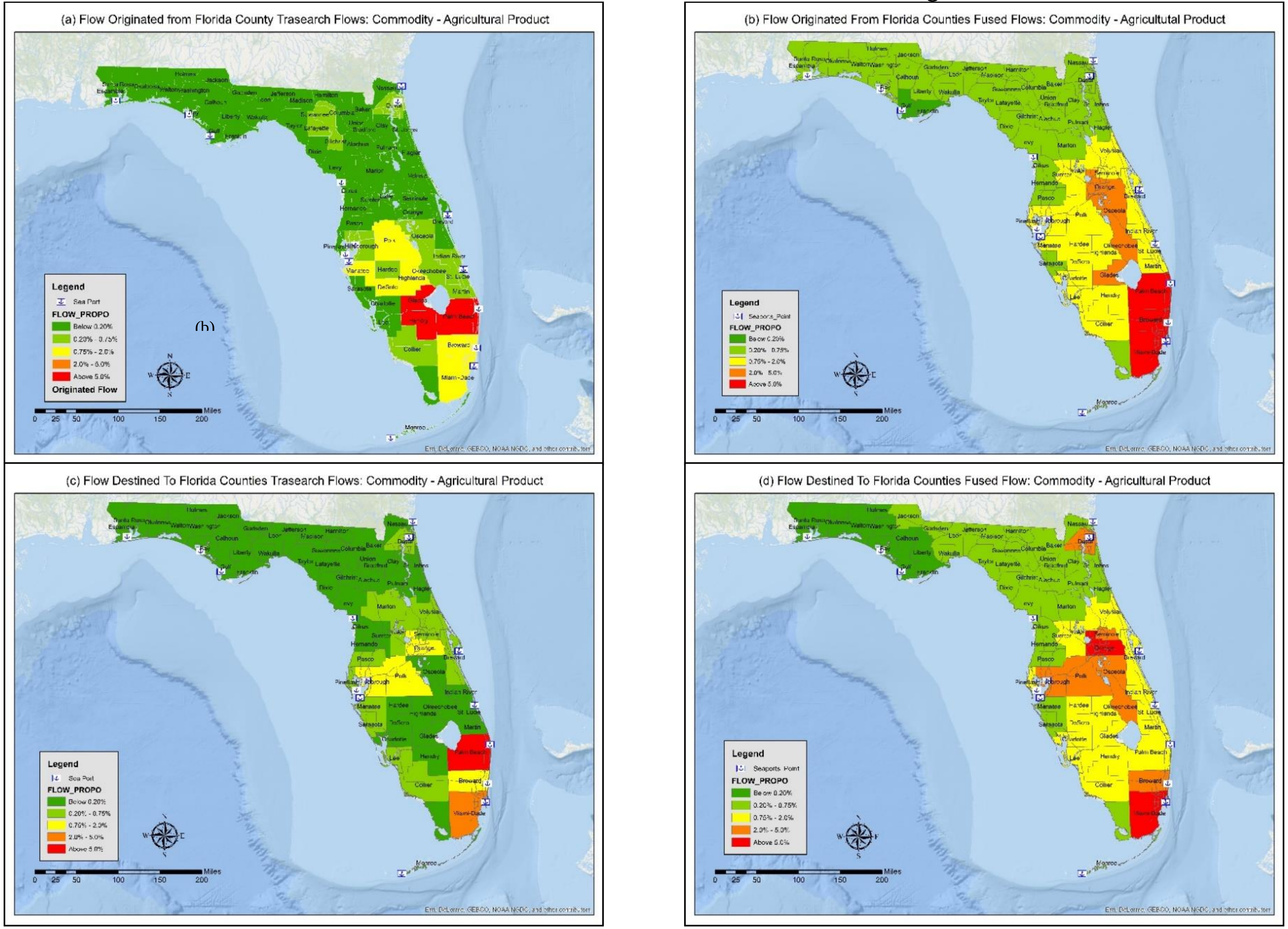


FIGURE 5 Transearch and Fused Flows within Florida Counties for Agricultural Product



(a) Flow Originated from Florida County Trasearch Flows: Commodity - Agricultural Product

(b) Flow Originated From Florida Counties Fused Flows: Commodity - Agricultutal Product

(c) Flow Destined To Florida Counties Trasearch Flows: Commodity - Agricultural Product

(d) Flow Destined To Florida Counties Fused Flow: Commodity - Agricultural Product

FIGURE 6 Transearch and Fused Flows within Florida Counties for Food

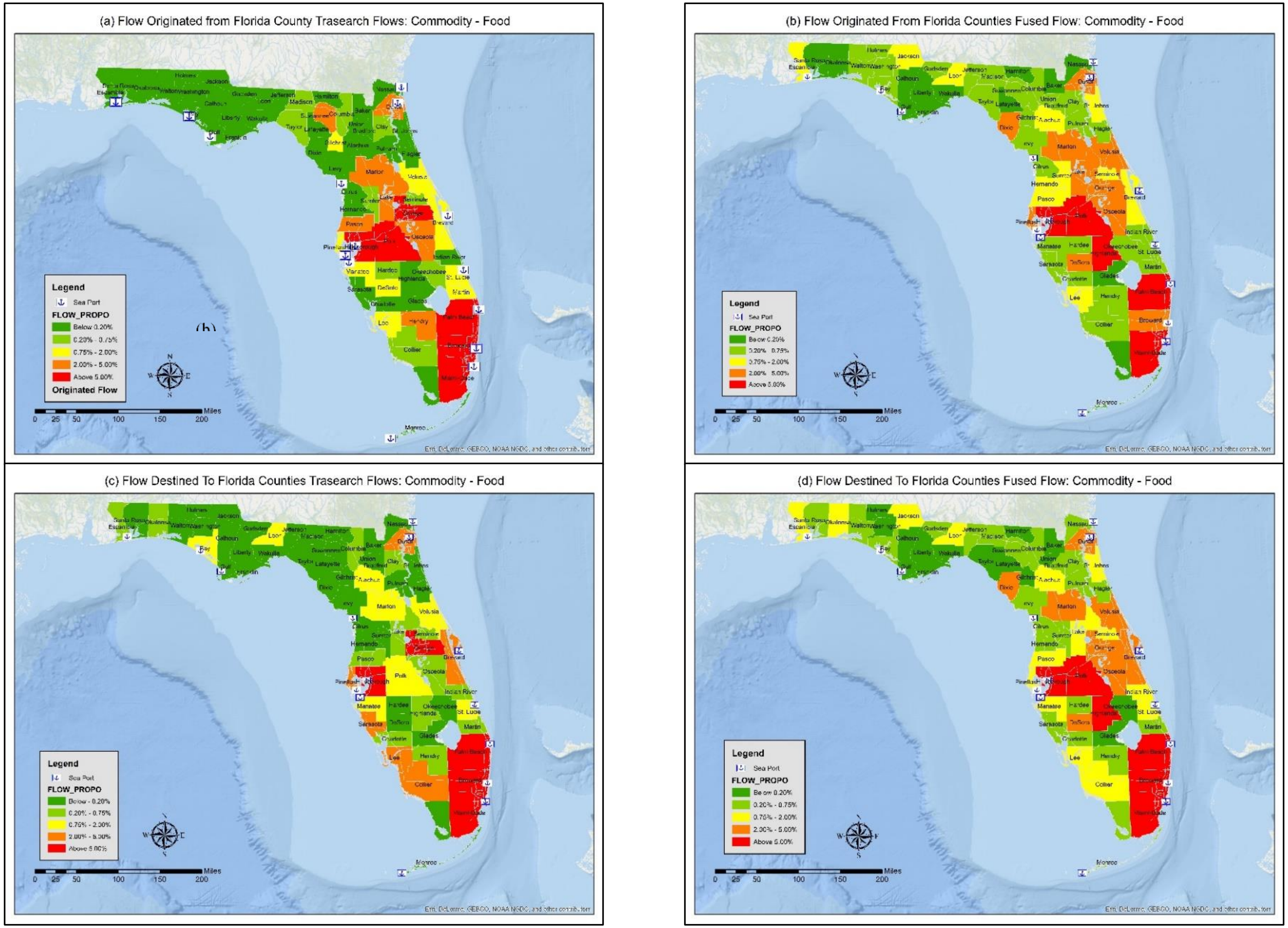
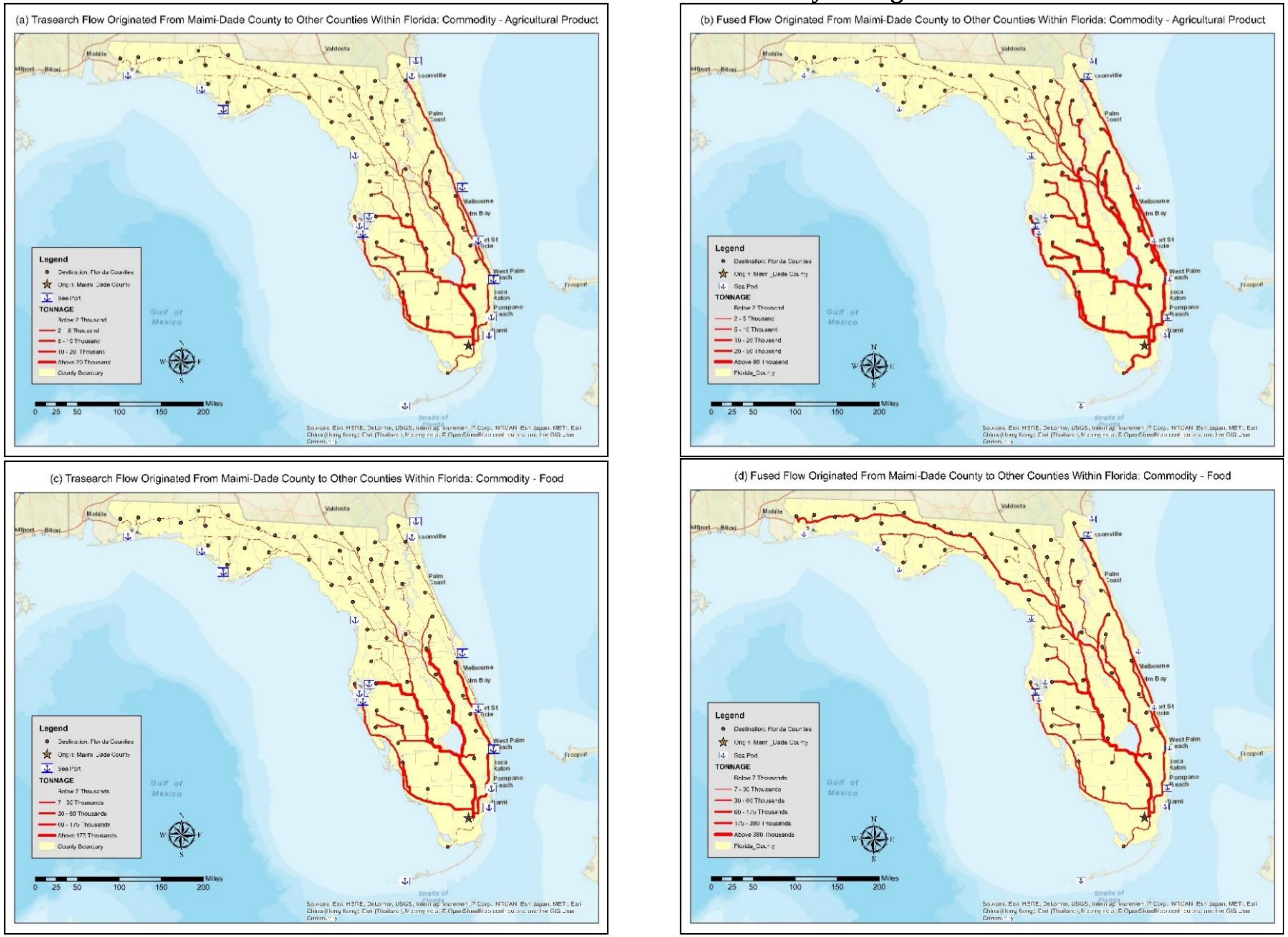

(a) Flow Originated from Florida County Trasearch Flows: Commodity - Food
Legend
Sea Port
FLOW_PROPO
Below 0.20%
0.20% - 0.75%
0.75% - 2.00%
2.00% - 5.00%
Above 5.00%
Originated Flow
(b)
Miles
0 25 50 100 150 200
(b) Flow Originated From Florida Counties Fused Flow: Commodity - Food
(c) Flow Destined To Florida Counties Trasearch Flows: Commodity - Food
(d) Flow Destined To Florida Counties Fused Flow: Commodity - Food
Esri, DeLorme, GEBCO, NOAA NGDC, and other contributors

FIGURE 7 Transearch and Fused Flows from Miami-Dade County for Agricultural Products and Food



(a) Trasearch Flow Originated From Maimi-Dade County to Other Counties Within Florida: Commodity - Agricultural Product

(b) Fused Flow Originated From Maimi-Dade County to Other Counties Within Florida: Commodity - Agricultural Product

(c) Trasearch Flow Originated From Maimi-Dade County to Other Counties Within Florida: Commodity - Food

(d) Fused Flow Originated From Maimi-Dade County to Other Counties Within Florida: Commodity - Food

TABLE 1 Review of Earlier Studies

Study	Geographic Region	Dataset(s) Used	Research Objective	Methodology	Variables Used
Giuliano et al., 2010	Los Angeles	CFS, IMPLAN, WISERTrade, WCUS, SCAG	Estimate link specific truck flows	Data integration; I-O model; gravity model; user optimal network assignment	Small area employment data
Bujanda et al., 2014	Texas	FAF^3, Transborder freight flow, Maritime flow	Estimate state level flows from FAF^3 (import and export flows)	ArcGIS spatial analysis; network assignment	-
Aly and Regan, 2014	California	FAF^2	Disaggregate FAF commodity flow at the county level	Proportional weighting for both origin and destination	Truck VMT
Opie et al., 2009	New Jersey	FAF^2, Transearch (for validation)	Disaggregate FAF commodity flow at the county level	Proportional weighting	Total land area occupied by port (import and export flows); for domestic flows: commodity-specific employment, truck VMT, total employment, population
Ranaiefar et al., 2013	California	FAF^3	Develop structural commodity generation model at the FAZ level	Structural equation model	Employment, number of establishments, population, agriculture related variables (farm acreages), manufacturing sector GDP, energy-related data (refinery capacity)
Mitra and Tolliver, 2009	North Dakota	FAF^2, truck count data (validation)	Disaggregate truck flows (productions and attractions)	Proportional weighting (production); I-O model (attraction) gravity model (internal flow)	Two-digit NAICS employment count
Vishwanathan et al., 2008	Florida	FAF^2, Transearch (output cross-check)	Disaggregate FAF commodity flow at the county level	Proportional weighting; linear regression	Total employment, population, two/three-digit NAICS employment count

Study	Geographic Region	Dataset(s) Used	Research Objective	Methodology	Variables Used
Ruan and Lin, 2010	Wisconsin	FAF^2, Transearch (validation)	Comparison of different data synthesis method for disaggregating FAF flows	Proportional weighting; direct regression; optimal disaggregation model	Employment by industry type, number of intermodal facilities
Ross et al., 2016	Georgia	FAF^3, CBP, Census data	Disaggregate FAF flows to county and TAZ level	Spatial regression; proportional weighting	Three-digit NAICS employment count, population, freight network density
Oliveira-Neto et al., 2012	USA	FAF^3, CFS (validation)	Disaggregate FAF flows at the county level; estimate ton-mile by mode	Log-linear regression; gravity model	Total employment payroll
Sorratini and Smith, 2000	Wisconsin	CFS, Transearch	Disaggregate truck flows at the TAZ level	I-O model	Employment
Lim et al., 2014	California	FAF^3, FAF^2, Transearch (validation)	Disaggregate FAF flows at the county level	Linear regression	Population, employment, farm acreage and crop sales

Note: Commodity Flow Survey (CFS), Freight Analysis Framework (FAF), Freight Analysis Zone (FAZ), Traffic Analysis Zone (TAZ), Waterborne Commerce of the US (WCUS), Southern California Association of Government (SCAG), Input-Output (I-O), North American Industry Classification System (NAICS), County Business Pattern (CBP), Input-Output (I-O)

TABLE 2 Freight Flows by Weight for Within Florida Flows reported in Transearch and FAF4

FCC	Transearch Flow (million tons)	FAF4 Flow (million tons)	Ratio (FAF4 flow/ Transearch flow)
Agricultural Products	17.130	34.257	2.00
Minerals	51.593	191.119	3.70
Food	12.210	29.284	2.40
Nondurable Manufacturing	0.855	5.087	5.95
Lumber	5.232	19.636	3.75
Chemicals	1.715	10.281	5.99
Paper	3.039	2.797	0.92
Petroleum	13.611	59.766	4.39
Other Durable Manufacturing	5.122	12.908	2.52
Clay and Stone	24.146	39.951	1.65
Waste	7.466	29.179	3.91
Miscellaneous Freight and Warehousing	51.132	14.544	0.28
Total	**193.253**	**448.811**	**2.32**

** There is no flow for the commodity Coal in the within Florida flow

TABLE 3 Model Estimates for Agricultural Product and Food

Model	**Explanatory Variables**	**Agricultural Product**		**Food**	
		Estimates	***t*-stats**	**Estimates**	***t*-stats**
Transearch Module	**Intercept**	3.7763	99.4770	1.4275	9.7080
	Dummy for Origin/Destination				
	Jacksonville Origin	-0.7353	-5.4510	-[1]	-
	Miami Origin	1.9115	10.5330	-	-
	Tampa Origin	-	-	0.8029	4.3680
	Orlando Origin	-	-	0.4654	3.0710
	Orlando Destination	0.7980	8.4710	-	-
	Miami Destination	-1.8459	-13.2560	-	-
	Destination County Attribute				
	Number of Warehouses	3.8474	20.3630	-	-
	Number of Ports	-	-	0.1649	4.0970
	Number of intermodal facilities	-0.1948	-5.4730	-	-
	Network Length (in KM)	-	-	1.2484	10.5550
	Origin County Attribute				
	Network Length (in KM)	-	-	1.6818	18.8570
	Interaction Variables				
	Origin County Employment (in 10^3) * Destination County Employment (in 10^3)	0.7266	9.6940	-	-
	Origin County Employment (in 10^3) /Destination County Population (in 10^6)	-	-	-0.8736	-9.7270
	Standard Error of the Estimate for Transearch	1.8413	87.7510	2.4667	62.8660
FAF Module	**Path Distance (in 10 Miles)**	-0.0248	-1.199	-1.0858	-1.0200
	Standard Error of the Estimate for FAF	2.1615	22.591	0.8101	8.3480
Number of observations		4070		2447	
Log-Likelihood of the model		-11496.773		-6850.817	

[1] Variable not found significant

TABLE 4 County Level Link Flow Prediction for Agricultural Product and Food

FCC	Description of Flow	Mean (Thousand Tons)	Std. Dev. (Thousand Tons)	Total (Million Tons)	No of Observations	FAF4 vs TS Ratio	Fused Link flows vs TS Ratio
Agricultural Products	TS County to County Flow	4.209	179.222	17.130	4070	2.000	1.445
	Estimated County Level Link Flow	5.514	22.105	24.752	4489		
Food	TS County to County Flow	4.990	35.063	12.210	2447	2.400	1.624
	Estimated County Level Link Flow	4.417	37.167	19.830	4489		

TABLE 5 County Level Percentage Point Differences between Transearch and Fused flows

FCC	By Origin County		By Destination County	
	Mean	Standard Deviation	Mean	Standard Deviation
Agricultural Products	2.39	7.62	2.29	2.74
Food	1.37	9.42	1.44	3.31