Published in the Journal of the Transportation Research Forum, February 2010 and Conference Proceedings from the Transportation Research Board 88th Annual Meeting, January 2009

Developing Freight Analysis Zones at a State Level: A Cluster Analysis Approach

Date Submitted: July 30, 2008

Word Count: 4,405 words + 12 figures & tables* 250 = 7,405 words

Dr. Gregory A. Harris, P.E
Director, Office for Freight, Logistics & Transportation
University of Alabama in Huntsville
Huntsville, AL 35899
Telephone number: (256) 824-6060
Fax Number: (256) 824-6970
Email: harrisg@uah.edu

Dr. Niles Schoening
Professor of Economics, College of Business Administration
University of Alabama in Huntsville
Huntsville, AL 35899
Telephone number: (256) 824-7314
Fax Number: (256) 824-6970
Email: schoenn@uah.edu

Dr. Phillip A. Farrington
Professor of Industrial Engineering
Department of Industrial and Systems Engineering and Engineering Management
University of Alabama in Huntsville
Huntsville, AL 35899
Telephone number: (256) 824-6568
Fax Number: (256) 824-6733
Email: paf@ise.uah.edu

Dr. James Swain
Professor of Industrial Engineering
Department of Industrial and Systems Engineering and Engineering Management
University of Alabama in Huntsville
Huntsville, AL 35899
Telephone number: (256) 824-6749
Fax Number: (256) 824-6733
Email: jswain@ise.uah.edu

Dr. Michael D. Anderson, P.E.
Associate Professor of Civil Engineering
Department of Civil and Environmental Engineering
University of Alabama in Huntsville
Huntsville, AL 35899
Telephone number: (256) 824-5028
Fax Number: (256) 824-6724
Email: mikea@cee.uah.edu

Nitin Sharma
Doctoral Student
Department of Industrial and System Engineering and Engineering Management
University of Alabama in Huntsville
Huntsville, AL 35899
Telephone number: (256) 824-6749
Fax Number: (256) 824-6733
Email: sharman@email.uah.edu

ABSTRACT

The ability to plan and forecast freight demand to support transportation infrastructure investment decisions is limited by the lack of available data at a level of detail that is meaningful to the transportation planner. This paper develops an initial methodology for developing Freight Analysis Zones (FAZs) at a sub-state level to facilitate use of the data from the Freight Analysis Framework 2 (FAF2) database and industry surveys. The FAF2 database is based upon the Commodity Flow Survey and is a comprehensive public freight knowledgebase. However, with 114 zones nationwide (most states have one or two zones), the ability of a state or Metropolitan Planning Organization transportation planner to use the data without significant disaggregation is limited. Currently, there is no consensus regarding the means to disaggregate the original FAF2 data. This paper addresses this problem by developing a systematic method for partitioning a state into meaningful zones that support effective freight transportation planning and analysis. The paper tests the application of FAZs to disaggregate freight data for use in a statewide by model through a case study in Alabama. The paper concludes that FAZs can be effectively used without degrading the quality of the forecasts.

INTRODUCTION

The Freight Analysis Framework 2 (FAF2) database is currently the most complete public freight data available, but it is aggregated at the national level and distributed between 114 origins and destinations, shown in Figure 1. In the FAF2 database Alabama has two designated zones, the Birmingham area and the remainder of Alabama. This high level of aggregation is not conducive to analyzing the effect of freight on the transportation infrastructure at the state or local level. As a result, in its current form this data has limited use for state or metropolitan planning organization level transportation planning. In 2006, the Federal Highway Administration funded four pilot projects to develop methods to disaggregate the FAF2 to the county level [*1*]. Disaggregation at the county level within Alabama would require the development of a 67 by 67 matrix by commodity and mode. However, in states such as Texas and Georgia, with significantly more counties, this could be a much more arduous task.

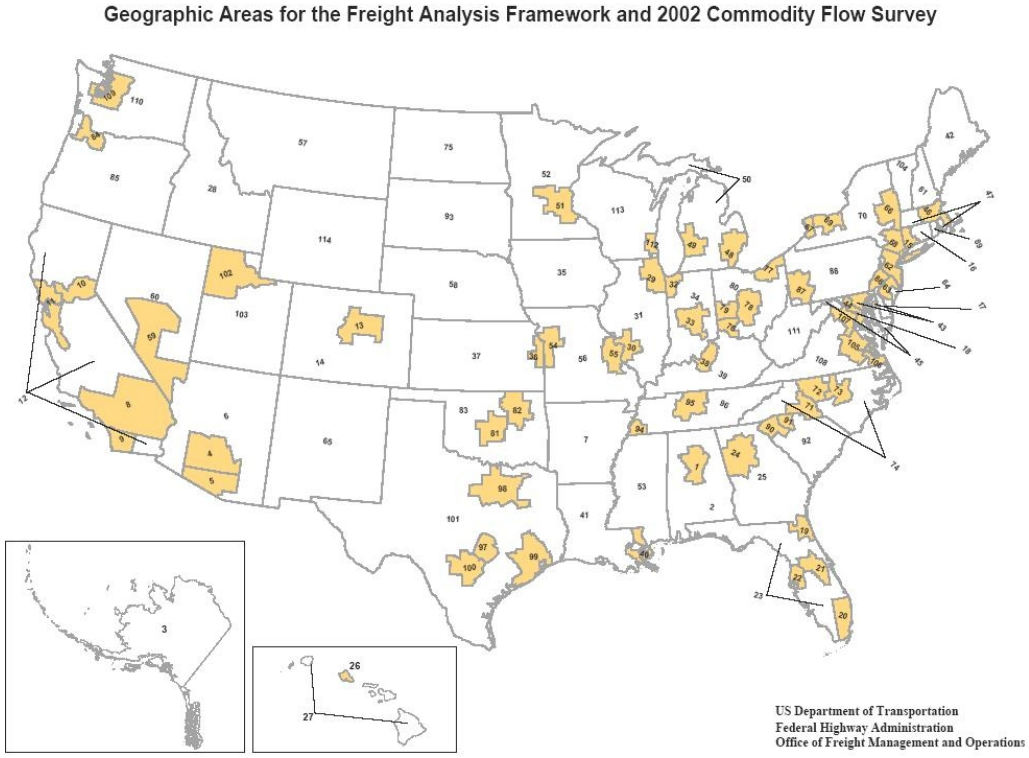

FIGURE 1 Geographic locations for FAF2 data [*2*].
(http://www.ops.fhwa.dot.gov/freight/freight_analysis/faf/cfs_faf_areas.htm**)**

In Alabama, ten counties contributed nearly 60% of total income and the top 20 accounted for three-fourths of all personal income in 2002 [*3*]. This situation is not significantly different from most other states. With resources for planning strained in most transportation budgets, effort applied to freight planning for areas where insignificant economic activity exists is not a responsible use of funds. It is theorized however, that areas of low economic activity could be aggregated into regions that contain enough economic activity to justify expending resources to plan for freight activity.

The decision to investigate the development of Freight Analysis Zones for use in freight planning and forecasting at a statewide level emerged from research into Freight Analysis Zones at the national level presented at the 2007 Transportation Research Board Annual Conference by Shin and Altman-Hall [*4*]. In their paper, Shin and Altman-Hall suggested that it would be beneficial to increase the number of Freight Analysis Zones (FAZs) in the FAF2 database from the current 114 total. Their paper described several methods to increase the number of zones, finally settling on approximately 400 utilizing the aggregation of zip codes to develop potential FAZs [*4*]. Although their purpose was to determine an optimal number of FAZs at the national level, the idea presented in the paper triggered the thought that the socio-economic factors being considered in a pilot study for the disaggregation of FAF2 data to the county level [*1*] could be used to aggregate counties into FAZs for freight planning and analysis at the sub-state level.

The goal of the FAF2 pilot program is to take national level data and disaggregate it to the county level from 114 national zones to 67 counties in Alabama and potentially 3500 counties in the U.S. With Alabama designated two zones in the FAF2 database, it is too aggregated to provide freight information for local or sub-state planning purposes. But, the research team at UAH believes that the county level may be too detailed for most states to use for freight planning. The authors believe that it is preferable for Alabama (and other states) to find a more "optimal" planning level that is, in the case of Alabama, "larger than 2 but less than 67." This optimal value should produce an aggregation that provides a necessary level of information without excessive detail.

This paper presents a cluster analysis based approach which strikes a middle ground and aggregates data from similar counties in close proximity. A guiding principle in the development of FAZs is that the zones should be homogeneous within the cluster but diverse from the surrounding clusters. Since the purpose of this initiative is to develop a methodology that transportation planners can use to enhance their freight planning, it is important that the final clusters promote the movement of traffic between clusters to provide the level of transactional data needed for planning purposes. The first step in the proposed approach uses economic data (employment, total value of shipments, personal income) and geographic data (longitude, latitude, and distance from interstate) to develop county clusters. The second step in the process validates and fine-tunes these clusters based on industry type and industry growth. The final step examines the results obtained when various levels of aggregation (county and FAZs) are used within a statewide freight flow model and concludes that there is not a significant reduction in accuracy obtained when properly applying the notion of FAZs.

BACKGROUND ON CLUSTER ANALYSIS

Cluster analysis is a multivariate technique that uses statistical procedures to form groups of entities called clusters based on certain pre-determined characteristics. More formally, a cluster is a collection of entities that have certain level of similarity or internal homogeneity between them and a distinct level of dissimilarity or external homogeneity with the entities forming other clusters. Two primary aspects under consideration for formation of clusters are: the type of similarity criteria and type of clustering method/technique.

The type of similarity criteria could be based on a certain type of distance measure or a concept that is common to all the entities across clusters. Thus, two entities could be a part of a cluster if they are within a certain geometric distance from each other or if they represent commonality with regard to a descriptive concept. This approach is similar to that used by

Moudon, et al. [*5*]for developing zones for metropolitan transportation planning. In this study characteristics and land use variables, such as density of activities, presence and agglomeration of destinations, block size, and transportation infrastructure attributes were used to determine the appropriate zones.

Distance measures such as Euclidean distance, Mahalanobis distance, Minkowski metric, Canberra metric, Czekanowski coefficient, Hamming distance etc. can be used to form distance-based clusters. Conceptual clustering uses formal definition of concepts generated by description languages along with the inherent structure of data to form clusters. COBWEB [*6*], CLUSTER/S [*7*] and LABYRINTH [*8*] are some examples of description languages that are used for concept definition. Due to stringent requirements related to formal definition of concepts and wider base of pre-requisite knowledge of the entities and attributes prior to clustering, conceptual methods are more difficult to implement and validate. For the current research, due to the complexity of the different economic attributes under consideration and the geographical zones, distance-based methods provide a simpler and effective foundation for cluster formation.

With a large number of entities and attributes associated with each entity, consideration of every single possibility/configuration becomes computationally expensive, thus calling for application of approximation methods or algorithms resulting into reasonable clusters. Such algorithms can be either hierarchical, resulting into a process that successively builds clusters through a series of partitions, or non-hierarchical involving identification of a seed as a central point and measuring distances from the same. Hierarchical methods can be either agglomerative by treating each entity as a cluster and iteratively combining entities to form clusters until a single cluster remains or divisive starting of by treating all entities as a single cluster and iteratively splitting entities to form clusters based on relative dissimilarities. As accurate determination of the initial seed in non-hierarchal methods can be cumbersome, and computationally expensive, hierarchical methods provide an efficient alternative for this research.

In hierarchical agglomerative methods, it is possible to form clusters on the basis of minimum distance (nearest neighbor or Single Linkage method), maximum distance (farthest neighbor or Complete Linkage method), average distance (Average Linkage method), minimum error sum of squares between clusters (Ward's method) and minimum distance between centroids of clusters (Centroid method). Ward's hierarchical clustering method proves effective when the intent is to minimize the loss of information associated with any iterative step in cluster formation. More formally, if the error sum of squares is represented by ESS_k for the k^{th} cluster then the total error sum of squares is given by $ESS_{total} = ESS_1 + ESS_2 + \ldots + ESS_k$. At any iteration, all possible combinations of entities are considered and the combination resulting into the smallest increase in the total error sum of squares is chosen for the union. This method is based on an assumption that clusters of multivariate observations are approximately elliptical in distribution. For the present research involving variables related to distance and economic parameters represented by different units and scales, controlling the loss of information per cluster formation and producing clusters of almost equal sizes are critical. Thus, Ward's method provides the necessary flexibility and setup to cater to the current problem statement. Statistical packages such as Minitab™ and Clustan™ provide efficient platforms for clustering algorithms and offer wide range of options for data display and graphical output providing useful and easy interpretation.

For hierarchical agglomerative procedures, Minitab™ provides the user with a wide range of options for linkage methods and distance measures for standardized and non-standardized variable formats for entities. It also gives the user the ability to control the final number of clusters and options for forming clusters based on either a distance measure or a similarity level. The matrix of distance between all pairs of cluster centroids and the cluster number for each entity can be stored separately based on user requirement. A graphical representation indicating the sequence of cluster formation relative to the distance measure or the similarity level, also known as a dendrogram can be plotted using Minitab™. The same functionalities are provided if cluster analysis is performed for attributes. Finally, a facility to perform the K-means method, a non-hierarchical procedure, is also available in Minitab™ for standardized and non-standardized variable formats. The general procedure in Minitab™ for cluster analysis on entities involves specification of the attributes that are required to be used in cluster formation along with the linkage method, distance measure, whether dataset has to be normalized and the number of clusters or desired similarity level. Options for storing different statistical metrics are also available.

Upon executing the routine, a dendrogram depicting the sequence of cluster formation is created. The console reveals useful information revealing the cluster number for each entity and the distance metric. Minitab™ thus provides useful features for statistical analysis required for a problem statement under consideration.

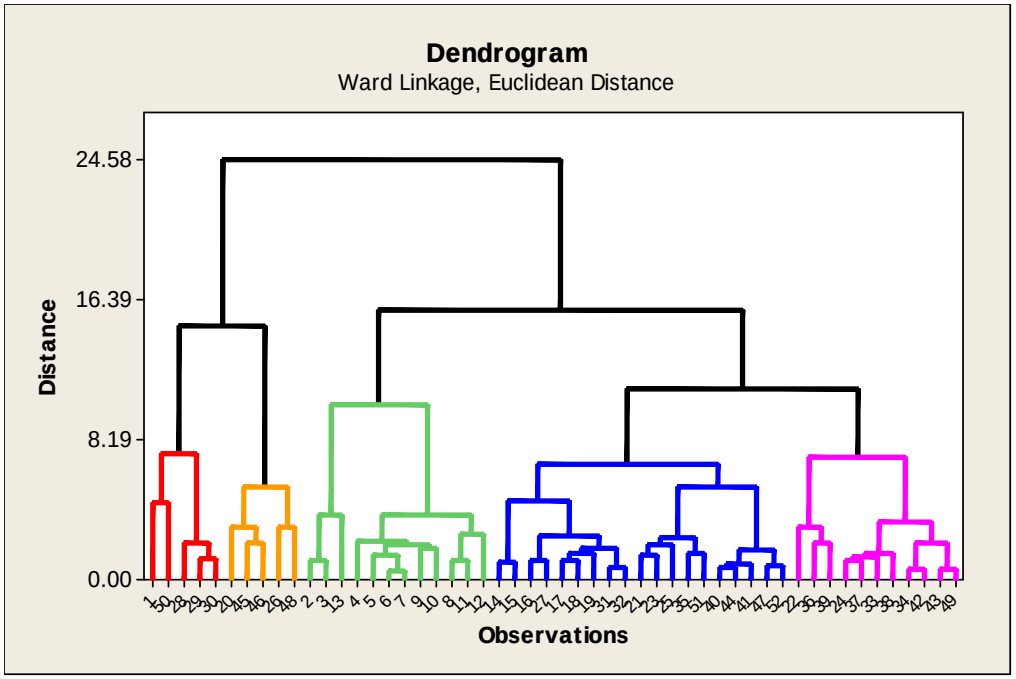

FIGURE 2 Sample Dendrogram from Minitab™.

DEVELOPMENT OF THE CLUSTERS

The process for the development of FAZs began with identification of the basic set of economic data that would be analyzed in order to define the analysis zones. Taking the counties as the basic unit of analysis, data was obtained on the employment level, payroll, value of shipments,

population, and personal income for each of the 67 counties in Alabama. This data set was then evaluated to form clusters using hierarchical clustering analysis. The clusters were formed using Wards method because it minimizes the within-cluster variance [*9*]. The distance between clusters considered for aggregation was measured using Euclidean distance.

This section presents an overview of the solution developed by the research team. In the quest to develop FAZs the researchers considered a variety of options but ultimately focused on clustering counties based on economic data and resulting in the development of eight potential solutions. All of the solutions utilized the economic data, however, in each of these cases the end result was several clusters that, while similar based on economic factors, were often widely dispersed geographically, a result that would not be conducive to effective freight planning and analysis. As a result, proximity measures were added to ensure that the location of the counties was taken into account in the development of the zones. Finally, the research team also noticed that the early outcomes seemed rather arbitrary; as a result it was felt there was a need to segment the state into regions to develop a more systematic way to grouping counties together. Therefore, the final solution builds clusters of counties within regions defined by the interstate highways that traverse Alabama.

One of the initial solutions investigated the formation of 11 clusters based on three variables, population, value of shipments, and personal income. It is clear from the solution shown in Figure 3 that without inclusion of proximity measures the clusters contain counties that are much more geographically dispersed. Thus, all future solutions included one or more measures of geographic proximity.

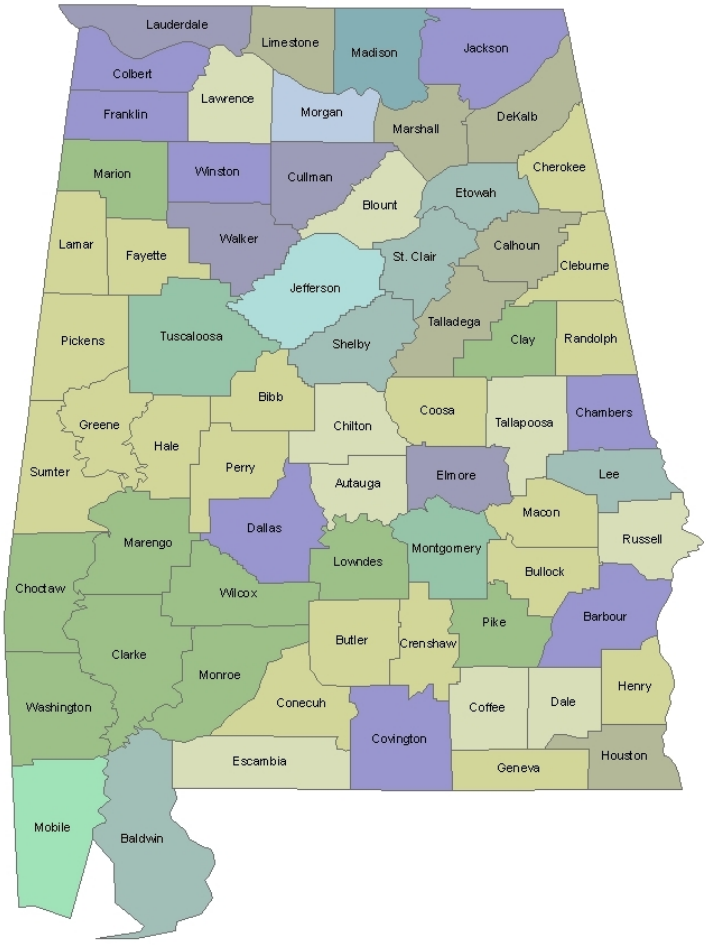

FIGURE 3 Cluster Solution of Counties Based Only on Economic Variables.

As stated above, the research team felt that regions of the state bounded by the interstates provided a more logical basis for defining sectors within the state. Figure 4 presents the layout of interstates that cross Alabama. For other states, the basis for sectors might be other transportation modes such as railroads or waterways. The use of interstates provided several attractive features because they provide natural boundaries and the objective was to pick up as much traffic flow on the interstate as possible and the most interstate traffic between zones to enhance the value of the data used in freight planning activities. Therefore, the UAH team chose to use interstate boundaries to divide the state into six planning sectors. Counties were allocated to sectors based on their proximity.

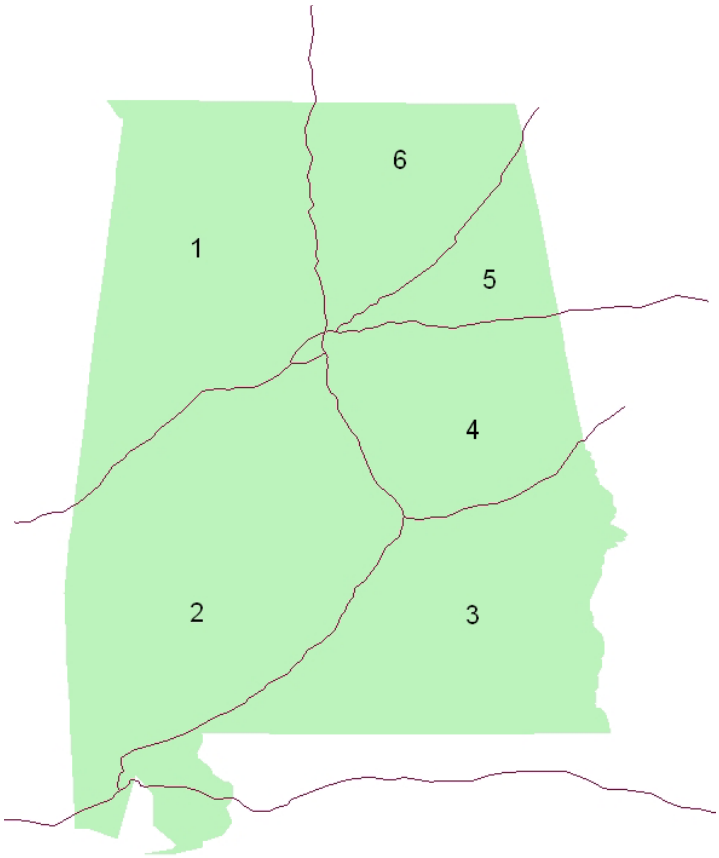

FIGURE 4 Interstate Based Sectors for Alabama.

Using the six sectors created by using the interstates as boundaries, a solution was generated based on a cluster analysis of counties within each interstate sector using economic variables as well as the county's longitude and latitude, and the distance from the interstate. This latter variable could be potentially important because it had been observed that counties closer to interstate appeared to have more freight traffic than counties further way from the interstate [*3*]. The solution shown in Figure 5 clustered counties within interstate sectors based on the economic variables, the proximity variables, and the distance of the county from the interstate. This resulted in 34 clusters. Review of the solution revealed that interstate sectors 3, 4, 5, and 6 contained too few counties for appropriate clustering. As a result, the research team decided to modify these interstate sectors by combining sectors 3 and 4 and sectors 5 and 6, resulting in a total of four interstate sectors. This modification resulted in fewer clusters but more homogenous clustering. The results are shown in Figure 6.

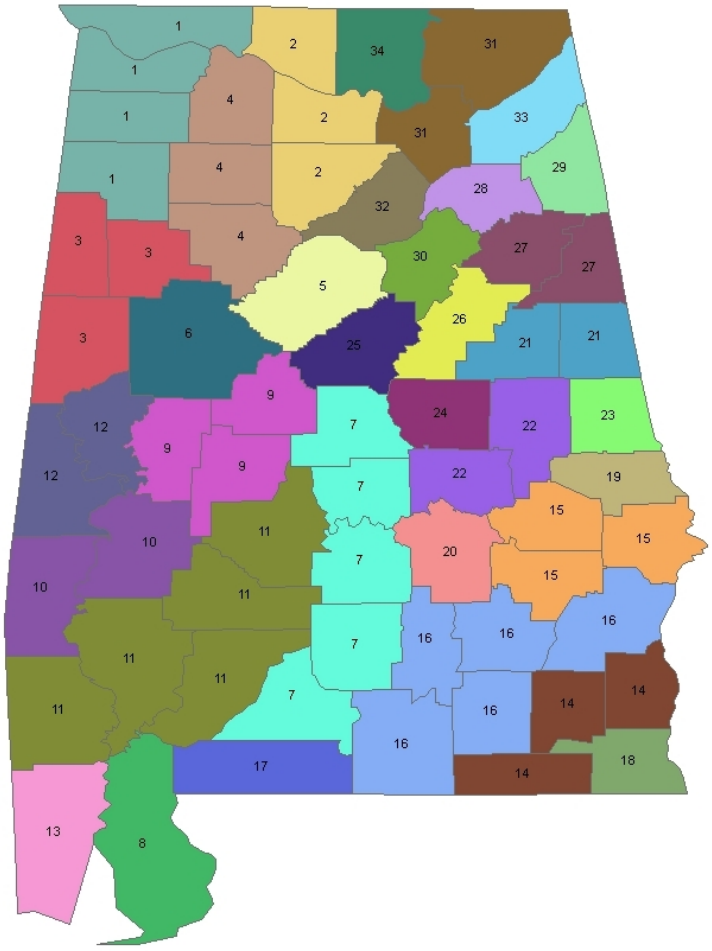

FIGURE 5 Cluster Solution within Interstate Sectors based on Economic Variables, Longitude, Latitude, and Distance from Interstate.

Figure 6 shows the cluster solution within the final four (4) interstate sectors based on the economic variables, proximity data, and each county's distance from the interstate. This approach resulted in a total of 27 clusters. The research team felt that this solution showed the most promise because the clusters were in close proximity within the natural boundaries provided by the interstates traversing Alabama.

After completion of the cluster analysis a refining step was added to the process. In this case, the 27 clusters were evaluated based on the type of industry and growth in each of the clusters. This step was performed in order to validate the defined clusters, and to refine the solution.

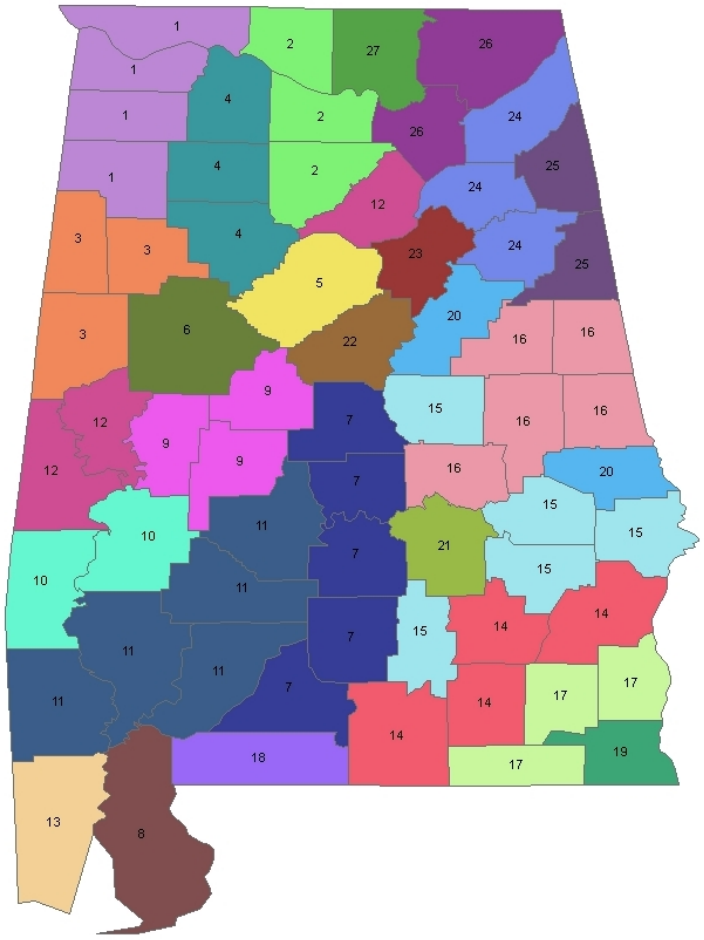

FIGURE 6 Cluster Solution within Modified Interstate Sectors based on Economic Variables, Longitude, Latitude, and Distance from Interstate.

Figure 7 shows the final cluster solution arrived at based on an evaluation of the solution shown in Figure 6 in which the individual clusters were refined based on types of industry and growth projections. The industries shown are the 17 largest industries in Alabama based upon employment [*3*]. Each industry listed employs more than 1000 people in the state.

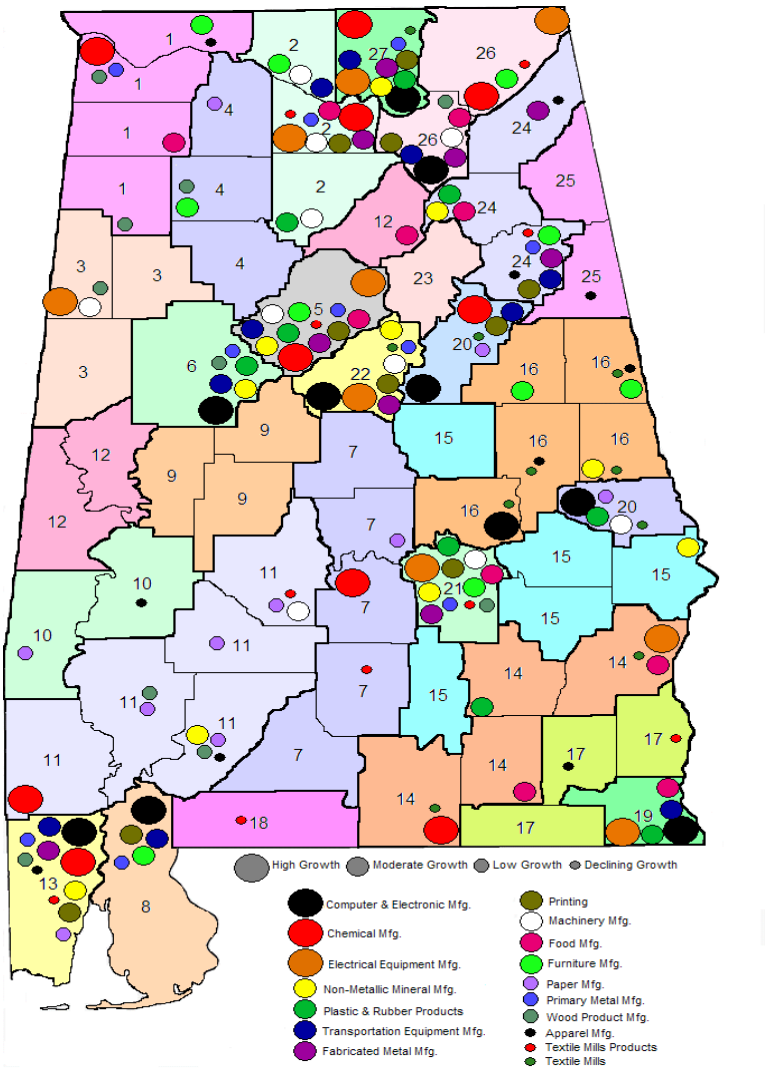

FIGURE 7 Final Cluster Solution with Industry Type and Growth Highlighted.

COMPARISION OF FINAL CLUSTER SOLUTION

To evaluate the differences between using Freight Analysis Zones (FAZ) versus all the counties in a state, a case study was developed using the State of Alabama Freight Model. The 67 county input file was created through a direct disaggregation of the FAF2 data using the various county proportions. Aggregating the various county data of the clustered zones that contributed to each FAZ created the 27 FAZ input file. The aggregated trips were then assigned to the county that best represented the economic center of the zone. This location became the origin or destination location for the FAZ. The Freight Distribution and Assignment Model was used to develop a truck trip exchange and determine the trucks forecasted to each section of roadway in the state. The distribution was performed using a gravity model on the truck production and attraction values. The gravity model develops a relationship between likely truck trip origins and likely destination based on a constrained value identifying where trips are required and determining the appropriate distance for the truck trips based on previously collected survey data. The

assignment of the truck trips was based on an All-or-Nothing procedure where all trips will take the shortest travel path from origin to destination. The shortest path was calculated as each segment of road in the model was attributed with segment distance and posted speed limit. The model operates in the TRANPLAN/CUBE® environment. The network contains almost 5,000 miles of roadway for Alabama and 15 roadways that serve as connections to surrounding states. The network is shown in Figure 8.

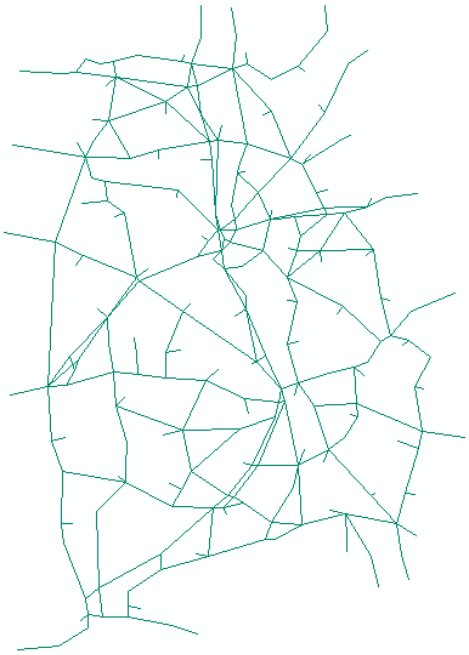

FIGURE 8 Network for the Alabama Distribution and Assignment Model.

To perform the case study, two input sets ofdisaggregated freight traffic were developed from the FAF2 database, one using counties and another using the 27 FAZs. Disaggregating the FAF2 database to either the county level or FAZ level developed the input files. The disaggregation of the FAF2 data included truck trips internal to Alabama, truck trips between Alabama and the other 49 states, and truck trips passing through Alabama. The disaggregation of the data was performed using a weighting of the economic factors, proportional to each county of FAZ contribution to Alabama's population, employment, personal income, and value of shipment. The two input files contained truck production and attraction values for either all 67 counties or for the 27 FAZ.After assigning the traffic to the network, the assignment can be reviewed visually for accuracy. See Figure 9.

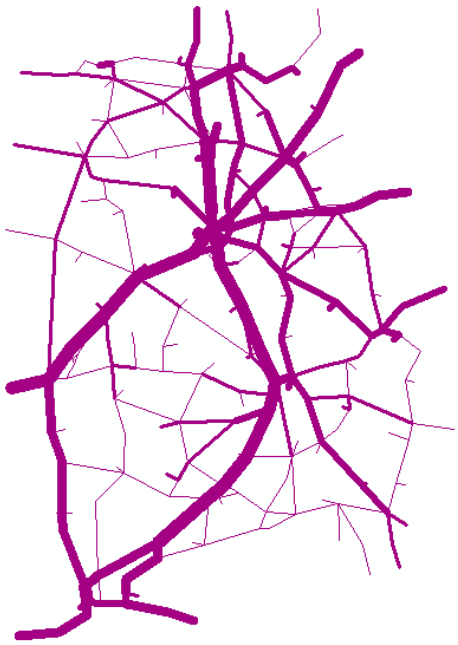

FIGURE 9 Assignment to the Network with Line Thickness Proportional to Assigned Volume.

To compare the performance of the two approaches (i.e., 67 counties versus 27 FAZs), a series of Alabama Department of Transportation (ALDOT) truck counts were added to the attributes for the network roadway segments. The ALDOT values for all roadway segments where the truck volume exceeded 1,000 trucks per day are identified in Figure 10.

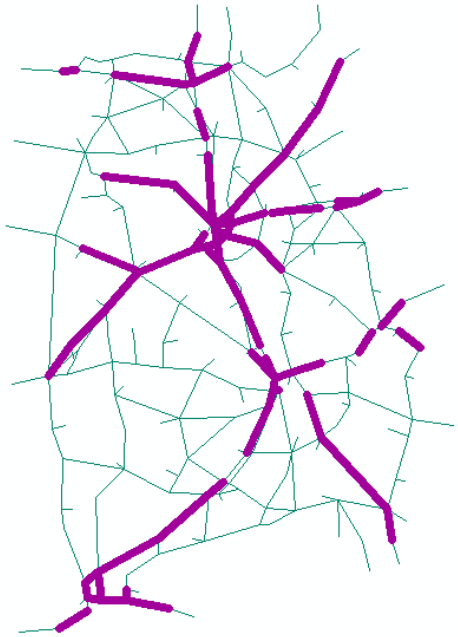

FIGURE 10 Location of ALDOT Truck Counts that Exceed 1,000 Trucks per Day.

Two scatter plots were developed to view the variations between the model assignment and the truck counts with both models assigned and the location of roadways where the daily truck volume exceeds 1,000 identified. Figures 11 and 12 show the scatter plots for the two models.

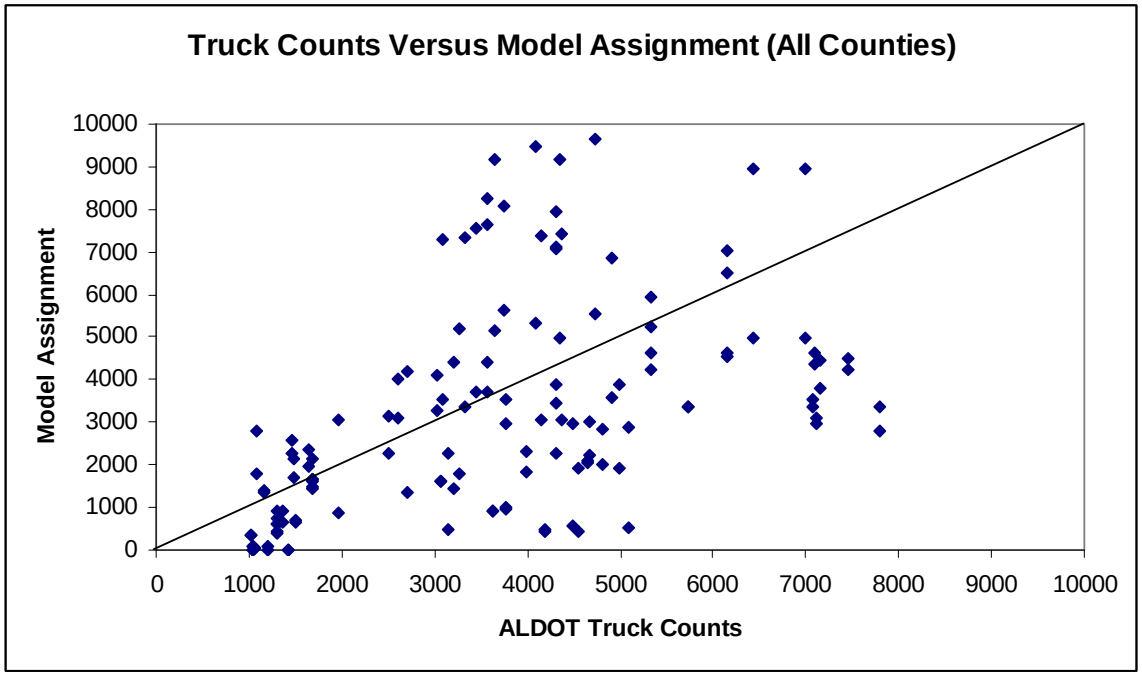

FIGURE 11 Scatter plot for the 67 County Model.

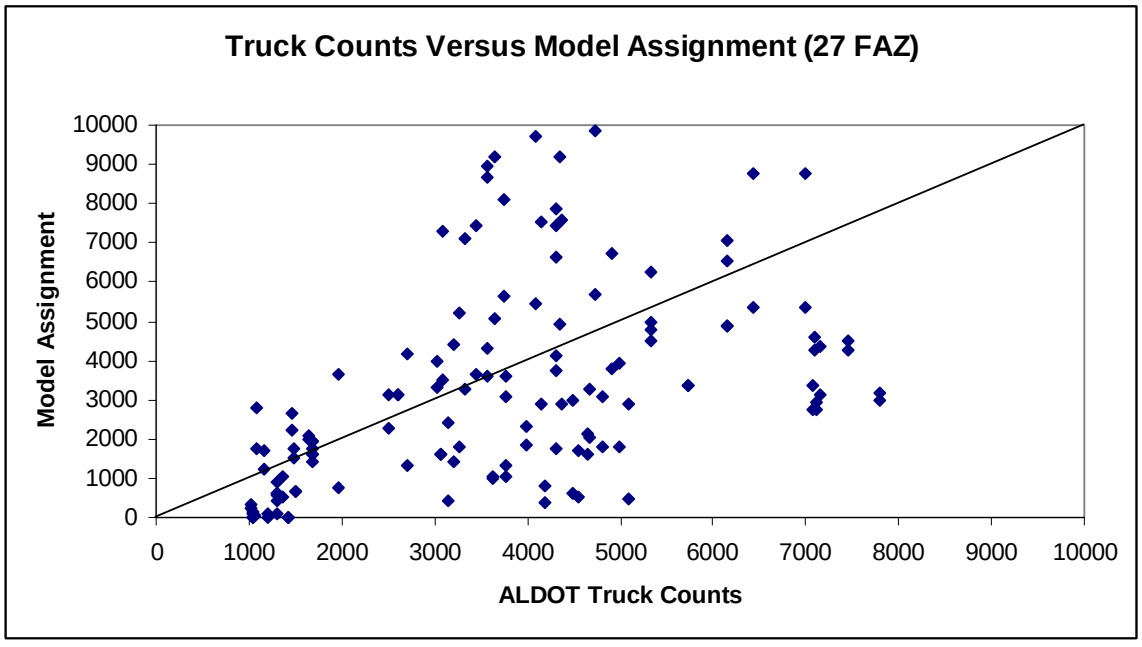

FIGURE 12 Scatter plot for the 27 FAZ Model.

To measure the difference between the model assignments using the two input levels (all 67 counties or the 27 FAZs), the Nash Sutcliffe's (NS) coefficient was employed [*10*]. The Nash-Sutcliffe value can range from -∞ to 1. An efficiency of 1 (E=1) corresponds to a perfect match of forecasted counts to the ground counts. An efficiency of 0 (E=0) indicates that the forecasted values are as accurate as the mean of the ground counts, whereas an efficiency less than zero ($-\infty<E<0$) occurs when the forecasted mean is less than the ground values. In other words, this coefficient gives us a measure of scatter variation from the 1:1 slope line of modeled truck counts vs. the ground counts. The more deviation of points from the slope line, the lower the coefficient. The greater the NS-value is the better the forecast. It can be calculated using the formula:

$$\text{NS-Coefficient} = 1 - \frac{\sum_1^n (ModeledCounts - GroundCounts)^2}{\sum_1^n (GroundCounts - MeanGoundCounts)^2}$$

The Nash Sutcliffe's statistic is considered the best measure of deviation between two data sets and used in many similar instances. Applying the Nash-Sutcliffe test for the two input files results in a NS-coefficient of 0.689 for the model that uses all 67 counties and a NS-coefficient of 0.679 for the model with 27 FAZ, indicating that there is no statistical difference in the assignments obtained using the 67 county model and the 27 FAZ model. This result supports the hypothesis that Freight Analysis Zones can be used to limit the data collection needs for freight planning without a reducing the quality of the assignment output.

CONCLUSIONS AND RECOMMENDATIONS

The ability to plan and forecast freight demand for transportation infrastructure is limited by the lack of available data at the level of detail that is meaningful to the transportation planner. The FAF2 database, based upon the Commodity Flow Survey, provides a publicly available freight knowledgebase for planning use. However, with 114 zones nationwide (and most states having two zones or less), the ability of the State or Metropolitan Planning Organization transportation planner to use the data is limited.

Disaggregation of the data to a more detailed level is needed to apply the freight flow data to whatever Statewide and Urban Planning model is currently being used. The fundamental problem is how to disaggregate the data to a usable level, without reducing the quality of the data to a point where its use would cause the introduction of excessive error. The initial use of counties as the disaggregation level for the freight data appeared promising and has easy initial understanding until the number of counties creates a data matrix that becomes excessively large and unwieldy. The research team believes that the ability to organize counties into Freight Analysis Zones provides a more efficient and effective way to organize the data into user-friendly form. The purpose of this paper was to develop an initial methodology for developing Freight Analysis Zones at a State level. The results found indicate that the development and use of Freight Analysis Zones for including freight in the overall transportation plan provides value and can improve the planning process.

Future research into the concepts of Freight Analysis Zones needs to continue through the examination of freight data disaggregation methods and travel model results. The various methodologies to disaggregate freight to the FAZs will help identify the impact of the using

these larger measurement units and the modeling of freight data will provide a mechanism to validate the various FAZs options.

ACKNOWLEDGEMENT
This research was sponsored in part by the U.S. Department of Transportation, Federal Transit Administration, Project No. AL-26-7262-00; The Federal Highway Administration, Project No. DTFH61-07-G-00007; and the Alabama Department of Transportation, Bureau of Research and Development, Research Project 930-682.

REFERENCES

1. Tang, T. FAF2 Pilot Project – Utilization of FAF2 Data By State and Local Governmental Agencies. Federal Highway Administration, February 28, 2006.

2. Freight Analysis Framework Documentation – November 7, 2007. http://www.ops.fhwa.dot.gov/freight/freight_analysis/faf/index.htm

3. Killingsworth, W.R., G.A. Harris, M.D. Anderson, M.J. Faulkner, A. Holden, J. Hutcheson, L.C. Jennings, S.A. Mondesir, M. Rahman, N.C. Schoening, R. Seetharam, J. Siniard, K. Stanley, J. Thompson, and A.D. Youngblood. *Transportation Infrastructure in Alabama: Meeting the Needs for Economic Growth.* The University of Alabama in Huntsville, Office for Economic Development, 2005 Report to the U.S. Department of Transportation, 2005. Grant No. DTTS59-03-G-00008.

4. Shin, H. and L. Aultman-Hall. Development of Nation-Wide Freight Analysis Zones. In the Proceedings *of the TRB 86th Annual Meeting Compendium of Papers CD-ROM*, Transportation Research Board Annual Conference, January 21 – 25, 2006. Washington, D.C.

5. Moudon, A.V., S.E. Kavage, J.E. Mabry, and D.W. Sohn. A Transportation-Efficient Land Use Mapping Index. In *Transportation Research Record: Journal of the Transportation Research Board, No.1902,* Transportation Research Board of the National Academies, Washington, D.C., 2005.

6. Fisher, D.H. Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning* 2:139-17, 1987.

7. Stepp, R. and R. Michalski. Conceptual Clustering of structured objects: A goal-oriented approach. *Artificial Intelligence* (28):43-69, 1986.

8. Thompson, K. and P. Langley. Concept Formation in Structured Domains. *Morgan Kaufman*: chapter 4, 1991.

9. Lattin, J., J.D. Carroll, and P.E. Green. *Analyzing Multivariate Data*. Brooks/Cole – Thomson Learning, Pacific Grove, CA, 2003.

10. Nash, J. E. and J.V. Sutcliffe. River flow forecasting through conceptual models. Part I: A discussion of principles. *Journal of Hydrology*. 10, 282-290, 1970.