Air Route Clustering for a Queuing Network Model of the National Airspace System

James DeArmon,[1] Christine Taylor,[2] Tudor Masek,[3] and Craig Wanke[4]
The MITRE Corporation, McLean, Virginia, 22102

A network queuing model of the National Airspace System has been developed to support research into a strategic air traffic flow management capability. One of the challenges in the execution of the model is the size of the network – the computing resources required when modeling the entire United States are immense. As a way to reduce the network size, we investigate route clustering, i.e., grouping similar routes to reduce the number of paths between two airports. Clustering routes comes at a cost: as the number of clusters falls, the with-in cluster variability rises, and the solution quality is diminished. A trade-off curve for solution quality vs. cluster variability is developed for a sample problem involving seven major airports.

I. Introduction/Background

A prototype capability for strategic air traffic flow management is undergoing research and development. The capability, called Flow Contingency Management (FCM), will supply automated decision support for what currently is a mostly manual process.[1] It is recognized that strategic decisions made with a 2- to 24-hour time horizon will likely improve air traffic flows in the National Airspace System (NAS) by averting large-scale traffic congestion due to weather. The Next-Generation Air Transportation System (NextGen) mid-term concept reflects the need for this type of capability.

Basic functionality has been developed for the prototype, including the representation of weather and traffic forecasts, and the integration of the two forecasts for predictions of significant impact. At the operative look-ahead times, there is significant uncertainty in the forecasts of both weather and traffic and, therefore, it is not appropriate to represent traffic at the level of individual flights. Rather, an aggregate model has been developed whereby traffic is represented as flows (an undifferentiated count of flights progressing in quarter-hour steps) in a queuing network. An initial formulation of such a model uses historical aircraft routings, one-day-prior filed flight routes, and "day-of" filed and predicted counts as input to a regression model to create the demand on a network of routes between airports. In the network, routes are represented by sequences of airspace sectors,[5] demand is expressed as the fraction per route of total flow between airports, and airports are represented as source and sink nodes. The queuing network model operates by associating air traffic demand with a sequence of sectors, and advancing time in quarter-hour increments. Sectors have a finite capacity, and flights may queue before transiting a sector, if demand would exceed capacity.

In prior work, it was found that clustering airports reduced the network size and complexity, as well as the model's run-time.[2] In this paper, we explore another means of reducing network size: route clustering, i.e., grouping of similar routes between airports. Assessing similarity of routes requires a similarity/difference measure and we propose the use of a specialized algorithm called "edit distance," appropriate for lexical string representation, i.e., the sequence of sectors in a route.

The paper is organized as follows. The next section describes the clustering algorithm: edit distance, similarity/difference assessment, and selection of a clustering method. Subsequent sections examine initial results, selection of a similarity threshold, and trading-off regression model error and resultant network size. A final section summarizes findings and suggests a next step in the analysis.

---

[1] Principal Simulation and Modeling Engineer, Mail Stop N430.
[2] Lead Simulation and Modeling Engineer, Mail Stop N450, AIAA Member.
[3] Senior Operations Research Analyst, Mail Stop N450.
[4] Senior Principal Simulation and Modeling Engineer, Mail Stop N450, AIAA Senior Member.
[5] A sector is a bounded airspace region under the control of a single air traffic controller or small team.

II. Route Clustering

In a recent paper describing the modeling of air traffic demand as input to the FCM network queuing model,[3] it was remarked that routes between two airports (or airport clusters), when represented as a sequence of sector transits, often exhibit notable similarity. For example, the following are two historical routes between Atlanta Hartsfield and Chicago O'Hare airports:

Route 1: ZTL38 ZTL37 ZID94 *ZID93* ZID76 *ZID78* ZID89 ZAU34 ZAU32
Route 2: ZTL38 ZTL37 ZID94 *ZID92* ZID76 ZID89 ZAU34 ZAU32

where the elements in the list, such as ZTL38, are sector identifiers and the three-character prefix in each is the code for an Air Route Traffic Control Center (ARTCC):

ZAU – Chicago ARTCC
ZID – Indianapolis ARTCC
ZTL – Atlanta ARTCC

The two sequences of sector transits are rather similar, differing only in the underlined and italicized sector identifiers.

A. Similarity/Difference Metric – Edit Distance

An obvious model simplification is to group the routes into subsets or clusters to create a smaller queuing network. Clustering algorithms rely on a distance function to assess similarities and differences among the elements to be clustered. For the application at hand, where the information is lexical, an "edit distance" is appropriate for determining similarity among elements.[4] Simply stated, the edit distance is the number of operations needed by a text editor to transform String 1 into String 2. Three operations are considered: deletion, insertion, and replacement. For example:

String 1: a b c d e
String 2: a b d e f

In this example, the edit distance is 2: (1) deletion of "c," and (2) insertion of "f." When comparing sector sequences, the atomic elements are not single letters as above; rather, the "alphabet" is the set of all sector identifiers. Using the original example of two routes:

Route 1: ZTL38 ZTL37 ZID94 *ZID93* ZID76 *ZID78* ZID89 ZAU34 ZAU32
Route 2: ZTL38 ZTL37 ZID94 *ZID92* ZID76 ZID89 ZAU34 ZAU32

The edit distance is 2: (1) replacement of ZID93 with ZID92, and (2) deletion of ZID78.

Since the number of sectors transited is variable (e.g., there are fewer sectors between Atlanta and Chicago than between Los Angeles and New York), we normalize the edit distance by dividing by the sum of the lengths of the two strings, and multiplying by 100, so the interpretation is "percent difference":

Similarity/Difference_Measure = 100 x Edit_Distance/(length of Route 1 + length of Route 2)

Given a similarity/difference measure for the application, a challenge is to select a threshold for declaring routes as similar or non-similar.

B. Choosing a Clustering Algorithm

With a similarity/difference measure defined, all route pairs for an origin/destination (O/D) can be so assessed, and clustered. It is necessary to select a clustering algorithm, as there are many available. Sample problems were constructed, and several algorithms were exercised: Ward's Method, Single Linkage, and Leader Algorithm. (See Ref. 5 for descriptions of the first two, and Ref. 6 for a description of the third). The Leader Algorithm gave the best results. The advice in the open literature is, when first considering a dataset for clustering, to employ several different clustering methods, and to look for intuitively-appealing results.[7,8]

The steps of the Leader Algorithm are as follows:

1) The first route is assigned as leader of the first cluster.
2) Examine each route in turn.
3) A route within the similarity threshold distance of a leader is assigned to that cluster.
4) A route outside the similarity threshold distance of all leaders becomes the leader of a newly-formed cluster.

III. Initial Results

Some experimentation resulted in reasonable separation between clusters, via visual assessment, when the similarity threshold was set to 20 percent. Figures 1, 2, and 3 below show the results for clustering of about 35 routes between Los Angeles International and New York's Kennedy International Airports during three days in August 2013. Figure 1 shows two different routes in the same cluster, plotted as a sequence of sectors (a plan view showing horizontal boundaries). The two routes are difficult to distinguish on the graphic, they're so similar that they agree on many of their set of transited sectors. Figure 2 shows two routes in different clusters; visually there is a significant difference in the paths. Figure 3 shows the four leaders which represent the entire data set; sector centroids (defined here as the average of the (x,y) values of the vertices of a sector at a specified altitude "slice") are connected via lines for clarity.

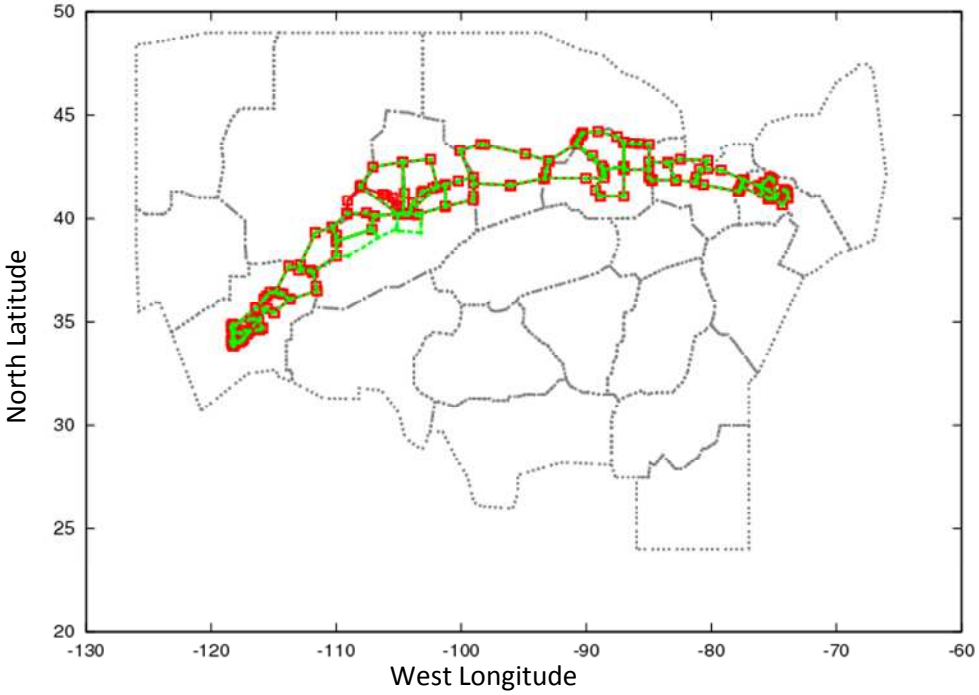

Figure 1. Two slightly differing routes in the same cluster, plotted as sequences of sectors; Route 1 is red lines and squares, Route 2 is green lines and dots. Area boundaries define U.S. ARTCCs.

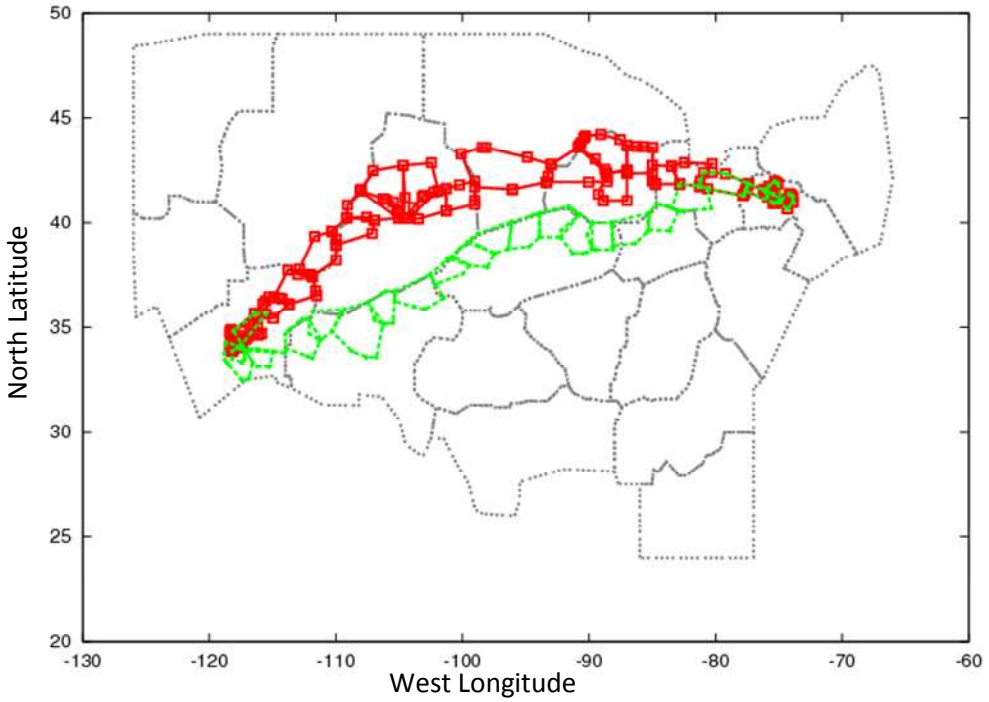

Figure 2. Two routes in different clusters, plotted as sequences of sectors; Route 1 is red lines and squares, Route 2 is green lines and dots. Area boundaries define U.S. ARTCCs.

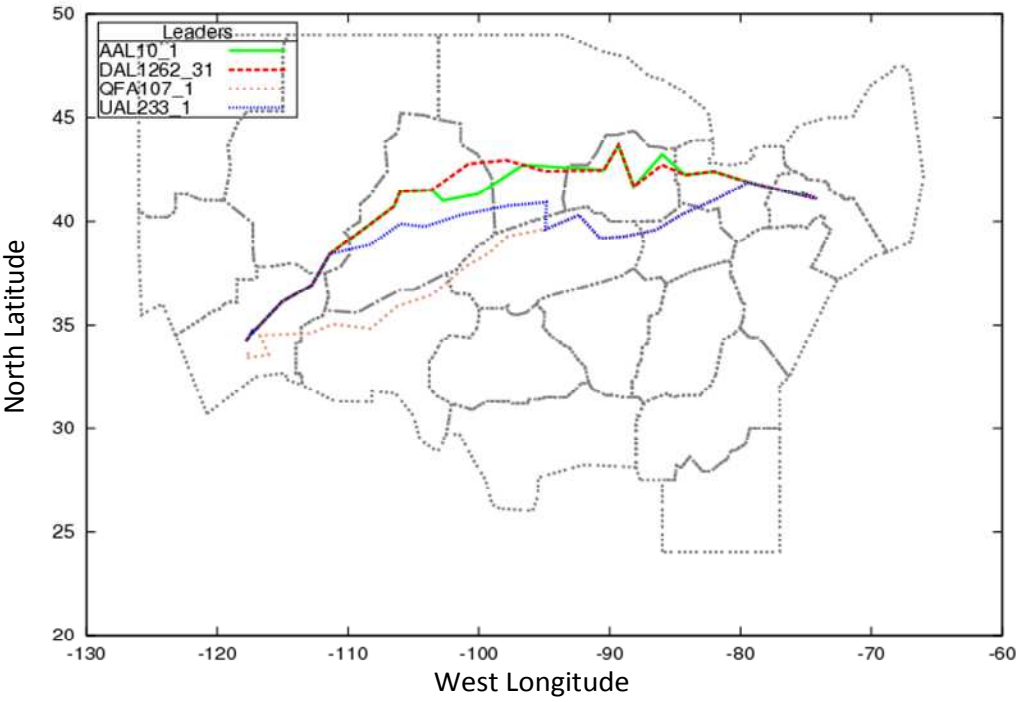

Figure 3. Four cluster leaders; routes are lines connecting cluster centroids. Area boundaries define U.S. ARTCCs

C. Improving the Representative Route of a Cluster

Notwithstanding the display of cluster leader routes in Fig. 3, a known deficiency of the Leader Algorithm (as noted in Ref. 6), is that the first observation encountered in the execution of the algorithm becomes the representative of the cluster. In some agglomerative clustering methods such as the K-means method, a cluster median or centroid observation is tracked, and becomes the representative of the cluster. As an improvement to the Leader Algorithm clustering used herein, a second-pass over the data was employed to identify a more central observation per cluster. For each cluster, a distance matrix was constructed using the pair-wise edit-distances; the route with the lowest total distance to all cluster members was identified, and deemed the best representative for the cluster. In Figure 4, one of the several clusters of routes from JFK to CLT is displayed in blue. The leader per the Leader Algorithm is displayed in red, while the centroid route, per the above logic, is displayed in green. A visual assessment suggests that the procedure of replacing the leader with the centroid improves the representativeness for the cluster.

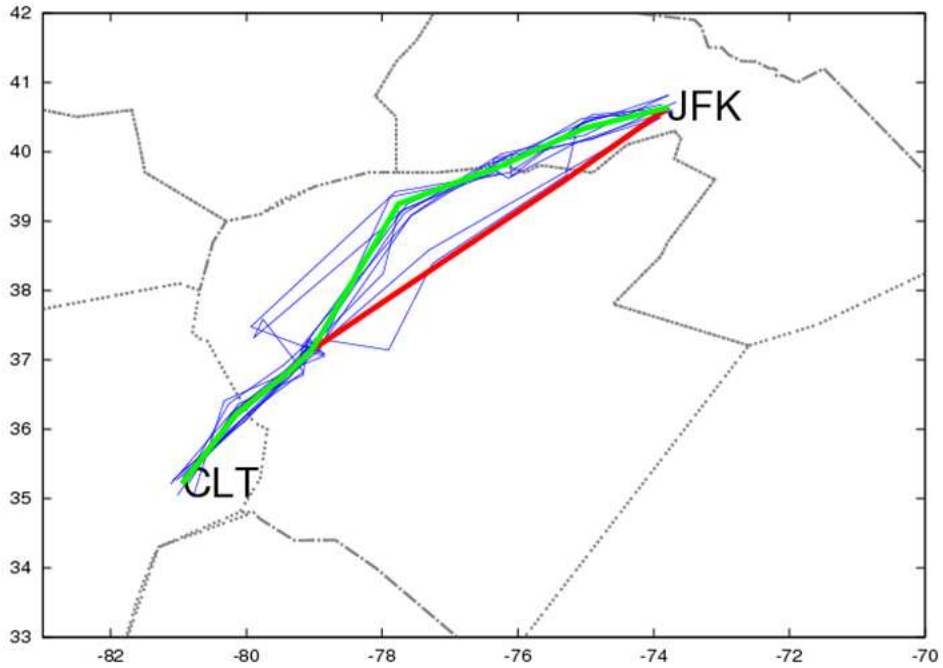

Figure 4. A route cluster (blue-colored routes) from JFK to CLT, with Cluster Leader (red) and Cluster Centroid (green). The x- and y- axes are north latitude and west longitude, respectively.

IV. Selecting a Similarity Threshold

The initial trials of the route clustering algorithm produced reasonable visual results. However, applying the clustering algorithm in practice, it would be necessary to select a similarity threshold quantitatively. Visual perusal yielded a threshold setting of 20% for what appear as reasonable differences between cluster leaders, and similarity between routes within a cluster, but testing and adjusting the threshold value in the context of the intended application was necessary.

To be useful, the route clustering approach should yield a reduced network size while not significantly diminishing the quality of the solution. As a quantitative basis of selecting a similarity threshold, results of a predictive regression model was used – as mentioned, one of the steps of FCM modeling is to predict, via regression analysis, the fraction of total flow per route, between airport pairs. (An operations model is used to generate demand counts[3], so that flow fractions can be converted to an integer number of flights per route.) A sample problem of flights to-and-from seven major airports in the NAS was selected for study. The airports are: Atlanta (ATL), Charlotte (CLT), Denver (DEN), Dallas-Ft. Worth (DFW), New York Kennedy (JFK), Los Angeles (LAX), and Chicago O'Hare (ORD). The resultant network has 42 city pairs (each of 7 airports has 6 destinations).

The predictive regression model was exercised on the seven airport dataset (using 60 days of prior flow fractions per route), and a comparison of actual to predicted was evaluated as "mean regression error," the units being average absolute difference in flow fractions for the multiple flows of an O/D pair. ("Actuals" are available when the model is being developed, using archived data.) As it happened, different O/D pairs achieved a minimum mean regression error at different similarity threshold values. A trade-off analysis was considered: reduced network size vis-à-vis maximum allowable similarity threshold (MAST) value.

In Fig. 5, the mean regression error vs. similarity threshold per airport pair (for a subset of pairs from the seven airport dataset, viz., the six destination-specific route sets with ATL as the origin airport) is shown. Here, the MAST value was set to 30%, and the minimum mean regression error was realized at varying levels of the similarity threshold as it was varied from 0% to 30% for these six O/D pairs. (Zero percent corresponds to the original set of routes without any clustering.)

A natural limit to the MAST value occurs at about 50%—at that point, some O/D pairs have all their routes aggregated into a single cluster, meaning the fraction of the flow being carried on the single-cluster route is 100%. An anomalous situation with regression modeling arises: the mean regression error in this case is zero, since there is a single observation and therefore zero degrees of freedom in the regression model and a "perfect" fit: the flow fraction is predicted to be 100% on a single, clustered route, and that agrees with the actuals.

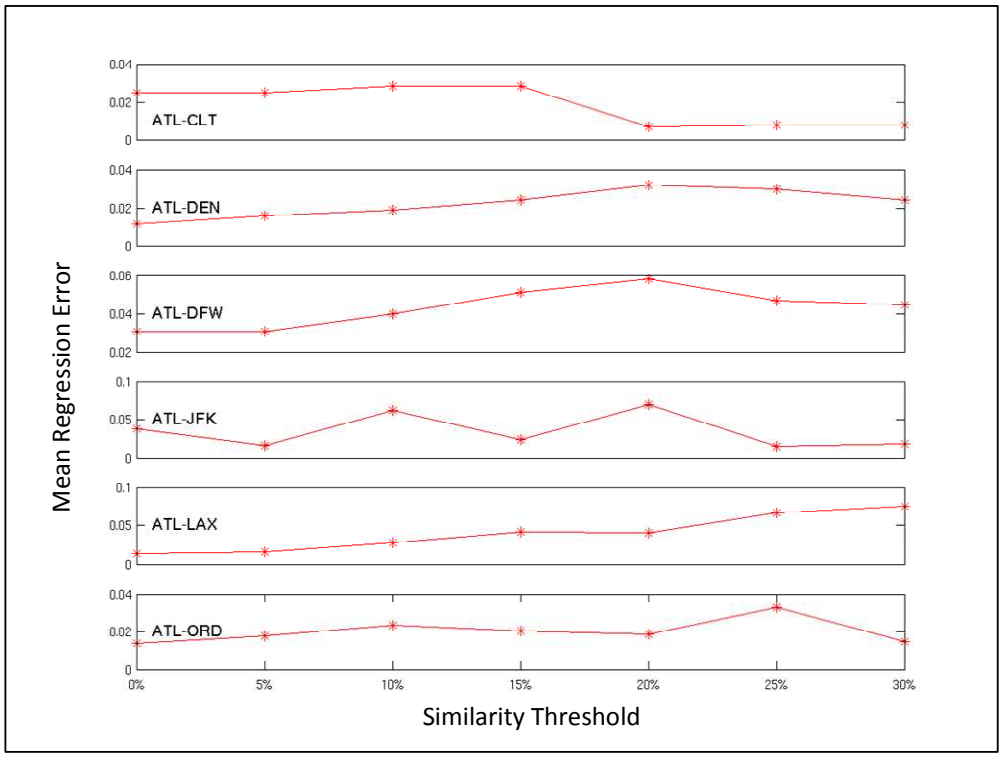

Figure 5. Similarity threshold vs. Mean Regression Error for selected airport pairs from the seven airport problem. Maximum Allowable Similarity Threshold (MAST) is 30%.

Multiple runs were undertaken, varying the MAST value from 0% to 50%. In each run, the 42 O/D pairs achieved their minimum regression error at a mix of similarity threshold values. As an example, a partial solution can be seen in Table 1 when the MAST value is set to 45%. Whereas the DEN-ATL routes achieved minimum mean regression error with no clustering at all (similarity threshold=0%), the DEN-LAX routes used the full MAST value of 45%.

Table 1. Similarity thresholds which yield minimum average regression error, for MAST of 45% (subset of results).

Origin-Destination	Similarity Threshold (%)
CLT-JFK	35
CLT-LAX	5
CLT-ORD	5
DEN-ATL	0
DEN-CLT	40
DEN-JFK	10
DEN-LAX	45

Since a composite solution is a mixture of OD-specific similarity threshold values, some overall cluster variability measure is needed in the trade-off analysis. That is to say, a trade-off between reduced network size and MAST would be over-simplifying. To that end, we sought overall cluster variability based on a measure of within-cluster variability.

A natural effect of clustering is that, for a given dataset, as the similarity threshold increases, the number of clusters diminishes, and the average within-cluster variability rises. Note that for some clustering algorithms, notably the well-known K-means approach, the within-cluster sum of squares can be computed directly (and used to help determine the proper number of clusters).[9] For the application at hand, however, K-means would not work well, since determining the difference or distance between two flight routes is not straightforward.

Earlier in this paper, we championed the edit-distance as a reasonable metric to compare two flight routes, if each is represented as a sequence of airspace sectors. Now for the within-cluster variability measure, it is necessary to consider actual flight path geometry to achieve a fair assessment of path variation. A reasonable measure of the difference between two routes is the maximum lateral deviation (MLD) between them. To define the flight path, we used the sequence of sector centroids. Figure 6 is an enlarged portion of Fig. 3, showing the MLD of two routes.

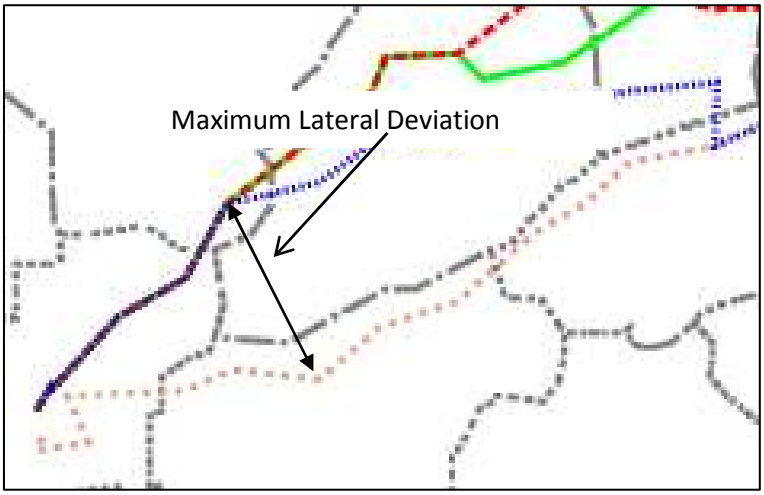

Figure 6. Maximum lateral deviation between two routes (Route 1 is pink dots, Route 2 is blue dots).

This approach[6] of determining path difference using MLD was successfully used in an earlier study comparing clear-weather routing to weather avoidance flight routes.[10] Here, the measurement is applied to all pairs of routes within a cluster, and the total deviation, in units of nautical miles, is then averaged. Another step in the computations is needed, since a composite solution contains 42 different O/D pairs, with greatly varying distances between origin and destination. This distance between origin and destination airports is an important factor in the extent of the MLD – it is more likely there is a large MLD for transcontinental flight routes between Los Angeles and New York than there is for short-hop flight routes between Charlotte and Atlanta, as there is a greater availability of airspace between the former pair. Therefore, the O/D distance is used as a divisor to normalize for this phenomenon. As a further refinement to this calculation, the average MLD is squared to magnify the contribution of large route differences.[7]

In summary, the computation of overall cluster variability for a composite solution is:

[(Average within-cluster MLD)^2 ÷ Distance(O,D)] averaged over 42 airport pairs.

The trade-off of reduced network size and overall cluster variability is shown in Fig. 7. A fitted curve (R^2 = 0.99) is overlaid on the point observations. At the upper left in the figure, no clustering (overall cluster variability = 0) is associated with a network size of about 2300 routes. At the lower right, corresponding to a MAST of 50% and the greatest overall cluster variability, the associated network size is about 1700. The trade-off curve is concave, meaning that a disproportionate increase in overall cluster variability (x-axis) is needed to effect a decrease in network size (y-axis). Such a trade-off curve will be constructed for the full problem, all flows in the entire NAS, and inform analysts as the research and development on the broader FCM project continues.

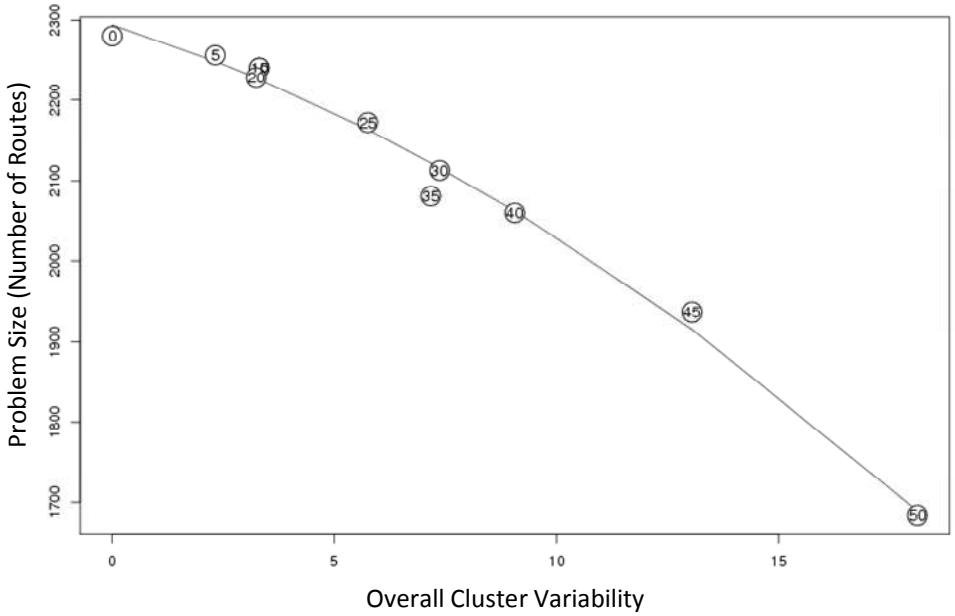

Figure 7. Network Problem Size vs. Average Cluster Variance. (Numbers in circles are MAST values. The (x,y) pairs for MAST=10 and 15 are coincident.)

[6] The approach is similar to one explored by MITRE colleague George Solomos. He produced a study of track variation measurements, but the study is not in the open literature.

[7] The mechanism is similar to the computation of the statistical variance of a population: squared deviations from the mean heighten the influence of outliers.

V. Summary

We have presented an approach for grouping or clustering of flight routes using the edit-distance metric applied to the sequence of transited airspace sectors. In the intended application, a network queuing model of the NAS, reducing the number of routes will reduce computer resources required to solve the problem. However, reducing the number of routes impacts the quality of a solution – in the case here, it increases the mean regression error in a model which predicts flows on routes. A trade-off analysis has been presented, for a limited seven-major-airport problem. In follow-on work this trade-off analysis will be pursued for a full flight route network representation of the NAS.

References

[1]Taylor, C. et al., "A Decision Support Tool for Flow Contingency Management," *AIAA Guidance, Navigation, and Control Conference*, Minneapolis, Minnesota, August 2012.

[2]Wang, L., Taylor, C., and Wanke, C., "An Airport Clustering Method for Air Traffic Flow Contingency Management," *Eleventh AIAA-ATIO Conference*, Virginia Beach, Virginia, September 2011.

[3]Wanke, C. et al., "Modeling Air Traffic Demand for a Real-Time Queuing Network Model of the National Airspace System," *Twelfth AIAA-ATIO Conference*, Indianapolis, Indiana, September 2012.

[4]Wikipedia article on Edit Distance, accessed November 2013: http://en.wikipedia.org/wiki/Edit_distance.

[5]Hastie, T, Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, 2nd ed., 2009, New York: Springer. pp. 520–528. ISBN 0-387-84857-6.

[6]Hartigan, J., *Clustering Algorithms*, Wiley and Sons (publisher), 1975, New York.

[7]Muller, E. et al., (2013) "Discovering Multiple Clustering Solutions: Grouping objects in different views of the data," *International Conference on Machine Learning* 2013, Carnegie Mellon University: http://www.cs.cmu.edu/~sguennem/publications/ICDE2012Tutorial.pdf

[8]Everitt, B. et al., *Cluster Analysis*, John Wiley & Sons, Ltd., 2011, West Sussex, UK.

[9]Everitt, B. and Hothorn, T., *A Handbook of Statistical Analyses Using R*. 2nd ed., 2010, Chapman and Hall/CRC, Boca Raton, Florida.

[10]DeArmon, J. et al., "An Estimation of the Benefits of Air Traffic Flow Management," *Eighth AIAA-ATIO Conference*, 2008, Anchorage, Alaska.

Disclaimer